

Lower Bounds on the Generalized Central Moments of the Optimal Alignments Score of Random Sequences

Ruoting Gong*

Christian Houdré†

Jüri Lember‡

November 24, 2016

Abstract

We present a general approach to the problem of determining tight asymptotic lower bounds for generalized central moments of the optimal alignment score of two independent sequences of i.i.d. random variables. At first, these are obtained under a main assumption for which sufficient conditions are provided. When the main assumption fails, we nevertheless develop a “uniform approximation” method leading to asymptotic lower bounds. Our general results are then applied to the length of the longest common subsequence of binary strings, in which case asymptotic lower bounds are obtained for the moments and the exponential moments of the optimal score. As a byproduct, a local upper bound on the rate function associated with the length of the longest common subsequences of two binary strings is also obtained.

AMS Mathematics Subject Classification 2010: 05A05, 60C05, 60F10.

Key words: Longest Common Subsequence, Optimal Alignment, Last Passage Percolation.

1 Introduction

Throughout this paper, $\mathbf{X}_n := (X_1, X_2, \dots, X_n)$ and $\mathbf{Y}_n := (Y_1, Y_2, \dots, Y_n)$ are two random strings, usually referred to as the (finite) sequences, so that any random variable X_i or Y_i , $i = 1, \dots, n$, takes its values in a fixed finite alphabet \mathcal{A} . The sequences \mathbf{X}_n and \mathbf{Y}_n are assumed to have the same distribution and to be independent. The sample space of \mathbf{X}_n and \mathbf{Y}_n will be denoted by \mathcal{X}_n ; clearly $\mathcal{X}_n \subseteq \mathcal{A}^n$, but, depending on the model, equality might not hold.

The problem of measuring the similarity of \mathbf{X}_n and \mathbf{Y}_n is central to many areas of applications including computational molecular biology (cf. [7], [10], [26], [28] and [32]) and computational linguistics (cf. [33], [23], [24] and [25]). In this paper we adopt the notation of [20], namely, we consider a general scoring scheme, where $S : \mathcal{A} \times \mathcal{A} \rightarrow \mathbb{R}^+$ is a symmetric *pairwise scoring function*, that assigns a score to each pair of letters from \mathcal{A} . Throughout, an *alignment* is a pair (π, μ) , where

*Department of Applied Mathematics, Illinois Institute of Technology, Chicago, IL 60616, U.S.A.
Email: rgong2@iit.edu.

†School of Mathematics, Georgia Institute of Technology, Atlanta, GA 30332, U.S.A.
Email: houdre@math.gatech.edu. Research supported in part by the grant # 246283 from the Simons Foundation and by a Simons Fellowship grant # 267336.

‡Institute of Mathematical Statistics, University of Tartu, Liivi 2-513 50409, Tartu, Estonia. Email: juril@ut.ee.
Research supported by Estonian Science foundation Grant no. 5822 and by institutional research funding IUT34-5.

$\boldsymbol{\pi} := (\pi_1, \pi_2, \dots, \pi_k)$ and $\boldsymbol{\mu} := (\mu_1, \mu_2, \dots, \mu_k)$ are two increasing sequences of positive integers such that $1 \leq \pi_1 < \pi_2 < \dots < \pi_k \leq n$ and $1 \leq \mu_1 < \mu_2 < \dots < \mu_k \leq n$. The integer k is the number of aligned letters, and $n - k$ is the number of *gaps* in the alignment. Note that our definition of gap differs from the one that is commonly used in the sequence alignment literature. More precisely, in most of the literature, a gap is a block of consecutive *indels* (insertion and deletion, formally denoted by one or more consecutive “–”) in both strings, and it depends on the alignment. For example, the left alignment below has four gaps, while the right one has five gaps.

$$\begin{array}{cccccccc} 1 & 1 & 3 & 1 & 2 & - & - & - & 3 & & 1 & 1 & 3 & 1 & 2 & - & 3 & - & - \\ 1 & - & 3 & - & 2 & 1 & 1 & 1 & - & & 1 & - & 3 & - & 2 & 1 & - & 1 & 1 \end{array} \quad .$$

Since we consider sequences of equal length, and since we do not have a gap opening penalty (which refers to a constant cost to open a gap of any length), our gap corresponds to a pair of indels, one on the \mathbf{X} -side and the other on the \mathbf{Y} -side. In other words, the number of gaps in this sense is the number of indels in either one of the sequences. Hence, in our framework, both the left and the right alignment above have three gaps.

Given a symmetric pairwise scoring function $S : \mathcal{A} \times \mathcal{A} \rightarrow \mathbb{R}^+$ and a *constant* gap price $\delta \in \mathbb{R}$, the score of the alignment $(\boldsymbol{\pi}, \boldsymbol{\mu})$, when aligning \mathbf{X}_n and \mathbf{Y}_n , is defined by

$$U_{(\boldsymbol{\pi}, \boldsymbol{\mu})}(\mathbf{X}_n; \mathbf{Y}_n) := \sum_{i=1}^k S(X_{\pi_i}, Y_{\mu_i}) + \delta(n - k).$$

In our general scoring scheme δ can be positive, although usually $\delta \leq 0$ penalizes matches with “–”. For negative δ , the quantity $-\delta$ is usually called the *gap penalty*. The optimal alignment score of \mathbf{X}_n and \mathbf{Y}_n is then defined as

$$L_n = L_n(\mathbf{X}_n; \mathbf{Y}_n) := \max_{(\boldsymbol{\pi}, \boldsymbol{\mu})} U_{(\boldsymbol{\pi}, \boldsymbol{\mu})}(\mathbf{X}_n; \mathbf{Y}_n),$$

where the maximum is taken over all possible alignments. To simplify notations, in what follows, we often set $\mathbf{Z}_n := (\mathbf{X}_n, \mathbf{Y}_n)$ so that $L_n = L_n(\mathbf{Z}_n)$. When $\delta = 0$ and the scoring function assigns the value one to every pair of common letters and zero to all other pairs, i.e.,

$$S(a, b) = \begin{cases} 1, & \text{if } a = b; \\ 0, & \text{if } a \neq b, \end{cases}$$

then $L_n(\mathbf{Z}_n)$ is just the maximal number of pairs of aligned letters – the length of the *longest common subsequences* (LCS) of \mathbf{X}_n and \mathbf{Y}_n , which is probably the most frequently used measure of global similarity.

1.1 A Summary of Known Results

The study of L_n is not only of theoretical interest but also has some practical consequences. For example, it is useful for distinguishing related (homologous) pairs of strings from unrelated ones. Unfortunately, for fixed n and an arbitrary distribution for \mathbf{X}_n and \mathbf{Y}_n , the distribution of L_n is unknown. In view of these, an alternative approach is to study the typical values and the

fluctuations of L_n . When \mathbf{X}_n and \mathbf{Y}_n are taken from an ergodic process, by Kingman's subadditive ergodic theorem, there exists a constant γ^* such that

$$\frac{L_n}{n} \rightarrow \gamma^* \quad \text{a.s. and in } L_1, \quad \text{as } n \rightarrow \infty. \quad (1.1)$$

In the LCS case, the existence of γ^* was first shown by Chvátal and Sankoff [8], but its exact value (or an identity for it) remains unknown even for independent i.i.d. Bernoulli sequences (although numerically estimated). Alexander [1] obtained the rate of convergence in (1.1) in the LCS case, which was extended to general scoring functions in [20], and to multiple (two or more) independent i.i.d. sequences with general score function in [12].

Since the exact distribution of L_n is rather hard to determine even for moderate n , it is natural to look for a limiting theorem, e.g.,

$$\frac{L_n - \mathbb{E}(L_n)}{n^\alpha} \xrightarrow{\mathfrak{D}} \mathcal{P}, \quad n \rightarrow \infty,$$

for some $\alpha \in (0, 1)$. Here \mathcal{P} is a limiting distribution, and $\xrightarrow{\mathfrak{D}}$ stands for convergence in law. Typically, one expects $\alpha = 1/2$, and \mathcal{P} to be a centered normal distribution, i.e.,

$$\frac{L_n - \mathbb{E}(L_n)}{\sqrt{n}} \xrightarrow{\mathfrak{D}} \mathcal{N}, \quad n \rightarrow \infty, \quad (1.2)$$

where \mathcal{N} stands for a centered normal distribution with variance $\sigma^2 > 0$. Under (1.2), for any $r > 0$, the r -th absolute moment of L_n would then be expected to grow at speed $n^{r/2}$, as $n \rightarrow \infty$. In particular, the variance would grow linearly, i.e.,

$$\lim_{n \rightarrow \infty} \frac{\text{Var}(L_n)}{n} = \sigma^2 > 0,$$

and then, (1.2) would be equivalent to

$$\frac{L_n - \mathbb{E}(L_n)}{\sqrt{\text{Var}(L_n)}} \xrightarrow{\mathfrak{D}} \mathcal{N}(0, 1), \quad n \rightarrow \infty. \quad (1.3)$$

Such a limiting theorem would allow to construct asymptotic tests, based on the optimal score L_n , hypothesizing that two given sequences \mathbf{X}_n and \mathbf{Y}_n are independent (not homologous) or not.

Note that in the gapless case, i.e., when $\delta = -\infty$, the optimal score is just the sum of the pairwise scores $L_n = \sum_{i=1}^n S(X_i, Y_i)$, and thus, under rather general assumptions on \mathbf{X}_n and \mathbf{Y}_n , the limiting theorem (1.3) (equivalently, (1.2)) holds true. This observation suggests to conjecture that (1.3) also holds in more general cases, in particular, in the LCS case. No such type of limiting theorem was known until [15], which proved (1.3) in the LCS case, when \mathbf{X}_n and \mathbf{Y}_n are independent i.i.d. random sequences such that $\text{Var}(L_n)$ admits a sublinear (in n) lower bound. The result in [15] upper-bounds the Monge-Kantorovich-Wasserstein distance, and in turn the Kolmogorov distance, implying weak convergence. The limiting theorem in [15] was extended to multiple independent i.i.d. sequences with a general score function in [12]. In contrast to [15], the result in [12] directly upper-bounds the Kolmogorov distance, which improves the rate of weak convergence, but still requires a (looser) sublinear variance lower bound. Hence, establishing a linear-order variance lower bound for a general distribution of X_1 implies immediately the corresponding limiting theorem (1.3).

Therefore, there is a direct connection between a variance lower bound and the limiting theorem (1.3), leading to an extra motivation to study the rate of convergence of $\text{Var}(L_n)$ as well as all the other central moments of L_n .

The study of the asymptotic order of $\text{Var}(L_n)$ in the case of LCS was first proposed by Chvátal and Sankoff [8], who conjectured that $\text{Var}(L_n) = o(n^{2/3})$, for \mathbf{X}_n and \mathbf{Y}_n independent i.i.d. symmetric Bernoulli sequences, based on Monte-Carlo simulations. By the Efron-Stein inequality, in the case of independent i.i.d. sequences with score functions satisfying (2.1) below, there exist a universal constant (independent of n) $C_2 > 0$, such that

$$\text{Var}(L_n) \leq C_2 n, \quad \text{for all } n \in \mathbb{N}. \quad (1.4)$$

(see Section 2 for more details). For the LCS case, this result was proved by Steele [29]. In [31], Waterman asked whether or not the linear bound on the variance can be improved, at least in the LCS case. His simulations showed that, in some special cases (including the LCS case), $\text{Var}(L_n)$ should grow linearly in n . These simulations suggest a linear lower bound on $\text{Var}(L_n)$, which would invalidate the conjecture of Chvátal and Sankoff. In the past ten years, the asymptotic behavior of $\text{Var}(L_n)$ has been investigated under various choices of sequences \mathbf{X}_n and \mathbf{Y}_n (cf. [5], [11], [17], [19], [21], [2], [3], etc). In particular, in [21], the asymptotic behavior of $\text{Var}(L_n)$ is investigated within two scenarios. In the first one, a general scoring function with a high gap penalty is considered, and \mathbf{X}_n and \mathbf{Y}_n are independent i.i.d sequences drawn from a finite alphabet (see also Theorem 6.2 and Remark 6.3 below). In the second one, the LCS case is studied when both \mathbf{X}_n and \mathbf{Y}_n are binary sequence having a multinomial block structure, which is a generalization of the model in [30] (the first analysis of the so-called *high entropy* case). In many ways, the present paper complements [21]. The results of [19] (the so-called *low entropy* case) were generalized in [16], and a linear variance lower bound was proved for an arbitrary finite alphabet (not just binary) and for the central moments of arbitrary order (not just the variance), but still under a strongly asymmetric distribution over the letters.

The goal of the present paper is to study in this general framework the lower bounds on generalized moments of the optimal alignment score L_n . To start with, in the next section we quickly show how to obtain upper moments and exponential bounds on L_n . In Section 3, we develop a general argument to find lower bounds for the generalized central moments of L_n . These lower bounds are obtained under a main assumption, and sufficient conditions for the validity of this main assumption are provided. Next, we relax some of these conditions (so that the main assumption is no longer satisfied), and develop a “uniform approximation” method giving rougher lower bounds for the generalized central moments of L_n . In Section 4, the general results are applied to the binary LCS case, and some refinements of those lower bounds are also presented. Finally, in Section 5, we obtain a lower bound on some moment generating function, which is then used to derive a (local) upper bound on the rate function associated with the binary LCS case.

2 Upper Bounds on Moments and Exponential Moments

In this short section, we briefly review the upper bounds on central moments and exponential moments of L_n . Recall that for any constant gap price $\delta \in \mathbb{R}$, changing the value of one of the $2n$

random variables X_1, \dots, X_n and Y_1, \dots, Y_n changes the value of L_n by at most K , where

$$K := \max_{u,v,w \in \mathcal{A}} |S(u,v) - S(u,w)|. \quad (2.1)$$

Thus, by Hoeffding's exponential martingale inequality, for every $t > 0$,

$$\mathbb{P}(|L_n - \mathbb{E}(L_n)| \geq t) \leq 2e^{-t^2/(nK^2)}, \quad (2.2)$$

which indicates that $V_n := (L_n - \mathbb{E}(L_n))/\sqrt{n}$ is subgaussian. Hence, for any $r > 0$,

$$\mathbb{E}(|V_n|^r) \leq r 2^{r/2} (9K^2)^{r/2} \Gamma\left(\frac{r}{2}\right) = r(18)^{r/2} \Gamma\left(\frac{r}{2}\right) K^r.$$

The above inequality provides an upper estimate, of the correct order, on the central absolute moments of L_n . Some further simple and direct computations allow to improve the constant. Let $r \geq 2$, $x > 0$, and let $\tilde{V}_n := |L_n - \mathbb{E}(L_n)|$. From (2.2),

$$\mathbb{E}(\tilde{V}_n^r) = \int_0^\infty \mathbb{P}(\tilde{V}_n \geq t^{1/r}) dt \leq x + 2 \int_x^\infty e^{-t^{2/r}/(nK^2)} dt.$$

Minimizing in x , i.e., taking $x = (K^2(\ln 2)n)^{r/2}$, and changing variables $u = t^{2/r}/(K^2n)$, lead to

$$\mathbb{E}(\tilde{V}_n^r) \leq K^r \left[(\ln 2)^{r/2} + r \int_{\ln 2}^\infty e^{-u} u^{r/2-1} du \right] n^{r/2} =: C(r) n^{r/2}.$$

When $x = 0$, the corresponding constant is slightly bigger than $C(r)$, and is given by

$$D(r) := rK^r \int_0^\infty e^{-u} u^{r/2-1} du = rK^r \Gamma\left(\frac{r}{2}\right).$$

Remark 2.1. For the LCS case, the following constant was obtained, as a consequence of a general tensorization inequality, in [16]:

$$E(r) := (r-1)^r 2^{r/2-1} \mathbb{P}(X_1 \neq Y_1).$$

For r close to 2, $E(r)$ is smaller than $C(r)$. Indeed,

$$E(2) = \mathbb{P}(X_1 \neq Y_1) \leq 1, \quad C(2) = 1 + \ln 2, \quad D(2) = 2.$$

However, for $r > 2$ large, $E(r)$ might be bigger, e.g.,

$$E(4) = 162 \mathbb{P}(X_1 \neq Y_1), \quad D(4) = 4\Gamma(2) = 4.$$

In addition to moment estimates, the above methodology also provides an upper bound on the moment generating function of $|V_n| = \tilde{V}_n/\sqrt{n} = |L_n - \mathbb{E}(L_n)|/\sqrt{n}$. Let $t > 0$ and $r > 1$. Then, with $K = 1$ and using (2.2),

$$\mathbb{P}(e^{t|V_n|} \geq r) = \mathbb{P}\left(\tilde{V}_n \geq \frac{\sqrt{n} \ln r}{t}\right) \leq 2e^{-(\ln r)^2/t^2}.$$

Thus,

$$\begin{aligned} M(t) &:= \mathbb{E}(e^{t|V_n|}) \leq 1 + 2 \int_1^\infty e^{-(\ln r)^2/t^2} dr = 1 + 2 \int_0^\infty \exp\left(-\frac{x^2}{t^2} + x\right) dx \\ &\leq 1 + t\sqrt{\pi} \left(1 + \operatorname{erf}\left(\frac{t}{2}\right)\right) e^{t^2/4} < 1 + 2t\sqrt{\pi} e^{t^2/4}, \end{aligned} \quad (2.3)$$

where $\operatorname{erf}(\cdot)$ denotes the Gaussian error function.

Remark 2.2. The upper bound on the moment generating function can clearly also be found through the upper bounds on the moments. Indeed, when $K = 1$ and with the bound

$$\mathbb{E}(|V_n|^r) \leq D(r) = r \Gamma\left(\frac{r}{2}\right) = 2 \Gamma\left(\frac{r+2}{2}\right) = 2\sqrt{\pi} \mathbb{E}\left(|\xi|^{r+1}\right), \quad r \in \mathbb{N},$$

where ξ is a $\mathcal{N}(0, 1/2)$ random variable, it follows that

$$\mathbb{E}\left(e^{t|V_n|}\right) = 1 + \sum_{r=1}^{\infty} \mathbb{E}(|V_n|^r) \frac{t^r}{r!} \leq 1 + 2\sqrt{\pi} \sum_{r=1}^{\infty} \mathbb{E}\left(|\xi|^{r+1}\right) \frac{t^r}{r!} = 1 + \frac{2\sqrt{\pi}}{t} \sum_{k=2}^{\infty} k \mathbb{E}\left(|\xi|^k\right) \frac{t^k}{k!}.$$

Since,

$$\mathbb{E}\left(|\xi| e^{t|\xi|}\right) = \mathbb{E}(|\xi|) + \frac{1}{t} \sum_{k=2}^{\infty} k \mathbb{E}\left(|\xi|^k\right) \frac{t^k}{k!} = \frac{1}{\sqrt{\pi}} + \frac{1}{t} \sum_{k=2}^{\infty} k \mathbb{E}\left(|\xi|^k\right) \frac{t^k}{k!},$$

while,

$$\mathbb{E}\left(|\xi| e^{t|\xi|}\right) = \frac{d}{dt} \mathbb{E}\left(e^{t|\xi|}\right) = \frac{d}{dt} \left[\left(1 + \operatorname{erf}\left(\frac{t}{2}\right)\right) e^{t^2/4} \right] = \frac{t}{2} \left(1 + \operatorname{erf}\left(\frac{t}{2}\right)\right) e^{t^2/4} + \frac{1}{\sqrt{\pi}},$$

the upper bound in (2.3) is recovered:

$$\mathbb{E}\left(e^{t|V_n|}\right) \leq 1 + 2\sqrt{\pi} \left(\mathbb{E}\left(|\xi| e^{t|\xi|}\right) - \mathbb{E}(|\xi|) \right) = 1 + t\sqrt{\pi} \left(1 + \operatorname{erf}\left(\frac{t}{2}\right)\right) e^{t^2/4}.$$

3 Lower Bounds on Generalized Moments

Let $\Phi : \mathbb{R}^+ \rightarrow \mathbb{R}^+$ be a convex, nondecreasing function. The main objective of this section is to find a lower bound for the expectation

$$\mathbb{E}\left(\Phi\left(|L_n(\mathbf{X}_n; \mathbf{Y}_n) - \mathbb{E}(L_n(\mathbf{X}_n; \mathbf{Y}_n))|\right)\right), \quad (3.1)$$

where, again

$$L_n = L_n(\mathbf{X}_n; \mathbf{Y}_n) = \max_{(\boldsymbol{\pi}, \boldsymbol{\mu})} U_{(\boldsymbol{\pi}, \boldsymbol{\mu})}(\mathbf{X}_n; \mathbf{Y}_n),$$

and where again the maximum is taken over all possible alignments.

Remark 3.1. In what follows, we deal with lower bounds on (3.1) with no further restrictions on Φ . This is somehow different from the upper bound case, where to the best of our knowledge, there is no uniform approach that applies to every convex nondecreasing Φ . Indeed, in the previous section, we established upper bounds for $\Phi(x) = |x|^r$, $r > 0$, and $\Phi(x) = e^{tx}$, $t > 0$. As shown in [16], Efron-Stein type inequalities can be applied to $\Phi(x) = |x|^r$, $r > 0$, via the Burkholder's inequality, giving the aforementioned constant $E(r)$. Another generalization of the Efron-Stein inequality, given, for example, in [6, Chapter 14], is the subadditivity inequality of the so called Φ -Entropy. However, it only applies to those functions Φ having strictly positive second derivative and such that $1/\Phi''$ is concave, e.g., $\Phi(x) = |x|^r$, for $r \in (1, 2]$, or $\Phi(x) = x \log x$. Moreover, the upper bounds on (3.1) for those Φ are linear in n as easily seen via, e.g., [6, Theorem 14.6], together with (2.1). In [6, Chapter 15], some further generalizations of Efron-stein inequality for $\Phi(x) = |x|^r$, $r \geq 2$, are also obtained, which, together with (2.1), also imply upper bounds of order $n^{r/2}$ on the r -th central moments (cf. Theorem 15.5 therein).

3.1 The Random Variable U_n and the Set \mathcal{U}_n

Recall that \mathcal{X}_n is the sample space of \mathbf{X}_n and \mathbf{Y}_n , so that $\mathcal{X}_n \times \mathcal{X}_n$ is the sample space of $\mathbf{Z}_n = (\mathbf{X}_n; \mathbf{Y}_n)$. In the sequel, we consider a function $\mathbf{u} : \mathcal{X}_n \times \mathcal{X}_n \rightarrow \mathbb{Z}$, so that $U_n := \mathbf{u}(\mathbf{Z}_n)$ is an integer-valued random variable. Let \mathcal{S}_n be the support of U_n , and for every $u \in \mathcal{S}_n$, let

$$\ell_n(u) := \mathbb{E}(L_n(\mathbf{Z}_n) | U_n = u),$$

and set $\mu_n := \mathbb{E}(L_n(\mathbf{Z}_n))$. Since $x \mapsto \Phi(|x - \mu_n|)$ is convex,

$$\mathbb{E}(\Phi(|L_n(\mathbf{Z}_n) - \mu_n|) | U_n) \geq \Phi(|\mathbb{E}(L_n(\mathbf{Z}_n) | U_n) - \mu_n|) = \Phi(|\ell_n(U_n) - \mu_n|),$$

so that

$$\mathbb{E}(\Phi(|L_n(\mathbf{Z}_n) - \mu_n|)) \geq \mathbb{E}(\Phi(|\ell_n(U_n) - \mu_n|)).$$

Let $\delta > 0$, and let the set $\mathcal{U}_n \subset \mathcal{S}_n$ be such that

$$\ell_n(u_2) - \ell_n(u_1) \geq \delta, \quad \text{for any } u_1, u_2 \in \mathcal{U}_n \text{ with } u_1 < u_2. \quad (3.2)$$

Since \mathcal{S}_n is finite, so is \mathcal{U}_n . Therefore, set

$$\mathcal{U}_n := \{u_1, \dots, u_m\} \quad \text{and} \quad k_0 := \max_{i=1, \dots, m-1} (u_{i+1} - u_i), \quad (3.3)$$

where $u_1 < u_2 < \dots < u_m$ and where $m = m(n)$ depends on n . Clearly, except perhaps in some very special cases, such a set \mathcal{U}_n formally always exists (it might even consists of two elements). As we will see, \mathcal{U}_n becomes useful if $\mathbb{P}(U_n \in \mathcal{U}_n) \geq c_0$, for δ and c_0 independent of n .

Lemma 3.2. *Under (3.2), for every $u_i, u_j \in \mathcal{U}_n$,*

$$|\ell_n(u_i) - \ell_n(u_j)| \geq \frac{\delta}{k_0} |u_i - u_j|.$$

Therefore, there exists $a_n \in [u_1, u_m]$, such that for every $u_i \in \mathcal{U}_n$,

$$|\ell_n(u_i) - \mu_n| \geq \frac{\delta}{k_0} |u_i - a_n|. \quad (3.4)$$

Proof. Let $g : [u_1, u_m] \rightarrow \mathbb{R}^+$ be a differentiable function, such that $g(u_i) = \ell_n(u_i)$, for every $i = 1, \dots, m$, and such that $g'(x) \geq \delta/k_0$, for every $x \in (u_1, u_m)$. Then, one of the following three mutually exclusive possibilities holds:

- $\mu_n < \ell_n(u_1)$,
- $\ell_n(u_1) \leq \mu_n \leq \ell_n(u_m)$,
- $\ell_n(u_m) < \mu_n$.

In the first case, for every $u_i \in \mathcal{U}_n$, $i = 1, \dots, m$,

$$|\ell_n(u_i) - \mu_n| = \ell_n(u_i) - \mu_n \geq \ell_n(u_i) - \ell_n(u_1) \geq \frac{\delta}{k_0} (u_i - u_1),$$

so (3.4) holds with $a_n = u_1$. The third case can be handled similarly with $a_n = u_m$. For the middle case, since g is increasing and continuous, there exists $a_n \in [u_1, u_m]$ such that $\mu_n = g(a_n)$. Hence, by the mean value theorem,

$$|\ell_n(u_i) - \mu_n| = |g(u_i) - g(a_n)| \geq \frac{\delta}{k_0} |u_i - a_n|,$$

completing the proof. \square

Now, in view of (3.4), since Φ is monotone and non-negative,

$$\mathbb{E}(\Phi(|\ell_n(U_n) - \mu_n|)) \geq \sum_{u_i \in \mathcal{U}_n} \Phi(|\ell_n(u_i) - \mu_n|) \mathbb{P}(U_n = u_i) \geq \sum_{u_i \in \mathcal{U}_n} \Phi\left(\frac{\delta}{k_0} |u_i - a_n|\right) \mathbb{P}(U_n = u_i),$$

i.e., the lower bound on the generalized moment is of the form

$$\mathbb{E}(\Phi(|L_n(\mathbf{Z}_n) - \mu_n|)) \geq \mathbb{E}(\Phi(|\ell_n(U_n) - \mu_n|)) \geq \sum_{u_i \in \mathcal{U}_n} \Phi\left(\frac{\delta}{k_0} |u_i - a_n|\right) \mathbb{P}(U_n = u_i) \quad (3.5)$$

$$= \mathbb{E}\left(\Phi\left(\frac{\delta}{k_0} |U_n - a_n|\right) \middle| U_n \in \mathcal{U}_n\right) \mathbb{P}(U_n \in \mathcal{U}_n). \quad (3.6)$$

When $\Phi(x) = |x|^r$, for some $r \geq 1$, then (3.6) becomes

$$\mathbb{E}(|L_n(\mathbf{Z}_n) - \mu_n|^r) \geq \left(\frac{\delta}{k_0}\right)^r \mathbb{E}(|U_n - a_n|^r | U_n \in \mathcal{U}_n) \mathbb{P}(U_n \in \mathcal{U}_n).$$

If $\mathbb{P}(U_n \in \mathcal{U}_n) \geq c_0 > 0$, with δ , k_0 and c_0 all independent of n , then

$$\mathbb{E}(|L_n(\mathbf{Z}_n) - \mu_n|^r) \geq \left(\frac{\delta}{k_0}\right)^r c_0 \mathbb{E}(|U_n - a_n|^r | U_n \in \mathcal{U}_n) \geq \left(\frac{\delta}{k_0}\right)^r c_0 \min_{a \in \mathbb{R}} \mathbb{E}(|U_n - a|^r | U_n \in \mathcal{U}_n).$$

Thus a lower bound on the centered absolute r th moment of $L_n(\mathbf{Z}_n)$ is obtained via conditioning on U_n . Typically, in applications, the random variable U_n counts the number of some particular letters in \mathbf{Z}_n , and, therefore, it has a binomial distribution. In this case, the right hand side of (3.6) is relatively easy to compute.

Remark 3.3. Note that the argument leading to (3.6) is independent of n . Therefore, (3.5) continues to hold whenever Φ depends on n , provided Φ_n remains convex and non-decreasing.

3.2 On the Existence of U_n and \mathcal{U}_n

As shown in the previous subsection, the lower bound on the Φ -moment can be obtained via the random variable U_n , provided that there exists a universal constant $\delta > 0$ such that the corresponding set \mathcal{U}_n has a large enough probability. In what follows, we show that the existence of a suitable \mathcal{U}_n can be guaranteed using a *random transformation*,

$$\mathcal{R} : \Omega \times \mathcal{X}_n \times \mathcal{X}_n \rightarrow \mathcal{X}_n \times \mathcal{X}_n,$$

such that, for most of the outcomes of Z , \mathcal{R} increases the score by at least some fixed amount $\varepsilon_0 > 0$. Here, with a slight abuse of notation, Ω is a sample space for the randomness involved in

\mathcal{R} . More precisely, the random transformation should be such that there exists a set $B_n \subset \mathcal{X}_n \times \mathcal{X}_n$ having a relatively large probability so that, for every $\mathbf{z}_n \in B_n$, the expected score of $\mathcal{R}(\mathbf{z}_n)$ exceeds the score of \mathbf{z}_n by ε_0 , namely,

$$\mathbb{E}(L_n(\mathcal{R}(\mathbf{z}_n))) \geq L_n(\mathbf{z}_n) + \varepsilon_0.$$

Above \mathbb{E} denotes the expectation over the randomness involved in \mathcal{R} (i.e., the integral over Ω). The requirement that the set B_n has a relatively large probability is formalized by requiring that

$$\Delta_n(\varepsilon_0) := \mathbb{P}(\mathbb{E}(L_n(\mathcal{R}(\mathbf{Z}_n)) - L_n(\mathbf{Z}_n) | \mathbf{Z}_n) < \varepsilon_0) \rightarrow 0, \quad \text{as } n \rightarrow \infty. \quad (3.7)$$

We will see that the faster $\Delta_n(\varepsilon_0)$ tends to zero, the better the lower bound is.

The random transformation should be associated with U_n and the set $\mathcal{U}_n = \{u_1, \dots, u_m\}$ so that when U_n takes its values in \mathcal{U}_n , then \mathcal{R} increases its value in \mathcal{U}_n and has no effect on the conditional distribution of \mathbf{Z}_n . To formalize this requirement, for every $u \in \mathcal{U}_n$, let $\mathbb{P}^{(u)}$ be the law of \mathbf{Z}_n given $U_n = u$, i.e.,

$$\mathbb{P}^{(u)}(A) = \mathbb{P}(\mathbf{Z}_n \in A | U_n = u), \quad A \subset \mathcal{X}_n \times \mathcal{X}_n.$$

Now, for every $u_i, u_{i+1} \in \mathcal{U}_n$, the following implication should hold: if $\mathbf{Z}_n^{(u_i)}$ is a random vector with law $\mathbb{P}^{(u_i)}$, then

$$\mathcal{R}(\mathbf{Z}_n^{(u_i)}) \sim \mathbb{P}^{(u_{i+1})}, \quad (3.8)$$

which means that for every $\mathbf{z}_n \in \mathcal{X}_n \times \mathcal{X}_n$,

$$\mathbb{P}(\mathcal{R}(\mathbf{Z}_n) = \mathbf{z}_n | U_n = u_i) = \mathbb{P}(\mathbf{Z}_n = \mathbf{z}_n | U_n = u_{i+1}).$$

In particular, (3.8) implies that if $\mathbf{u}(\mathbf{Z}_n) = u_i$, then $\mathbf{u}(\mathcal{R}(\mathbf{Z}_n)) = u_{i+1}$, but the converse statement might not be true.

Fast Convergence Condition. Our first general lower bound theorem assumes the existence of ε_0 so that $\Delta_n(\varepsilon_0)$ converges to zero sufficiently fast. To simplify the notation, in what follows, let $\tilde{\mathbf{Z}}_n := \mathcal{R}(\mathbf{Z}_n)$. The following theorem is a generalization of [21, Theorem 2.2].

Theorem 3.4. *For $n \in \mathbb{N}$, let there exist a random variable $U_n = \mathbf{u}(\mathbf{Z}_n)$, a set $\mathcal{U}_n = \{u_1, \dots, u_{m(n)}\}$, a random transformation \mathcal{R} satisfying (3.8), such that the following two assumptions hold:*

(i) *There exists a universal constant $A > 0$ such that*

$$L_n(\tilde{\mathbf{Z}}_n) - L_n(\mathbf{Z}_n) \geq -A. \quad (3.9)$$

(ii) *There exists a universal constant $\varepsilon_0 > 0$ such that*

$$\Delta_n(\varepsilon_0) = o(\varphi(n)), \quad n \rightarrow \infty, \quad (3.10)$$

where $\Delta_n(\varepsilon_0)$ is given by (3.7), and where φ be a function on \mathbb{N} such that

$$\mathbb{P}(U_n = u) \geq \varphi(n), \quad \text{for all } u \in \mathcal{U}_n. \quad (3.11)$$

Then, there exists $a_n \in [u_1, u_m]$ so that for n large enough,

$$\mathbb{E}(\Phi(|L_n(\mathbf{Z}_n) - \mu_n|)) \geq \sum_{u_i \in \mathcal{U}_n} \Phi\left(\frac{\delta}{k_0} |u_i - a_n|\right) \mathbb{P}(U_n = u_i) \geq \sum_{u_i \in \mathcal{U}_n} \Phi\left(\frac{\delta}{k_0} |u_i - a_n|\right) \varphi(n), \quad (3.12)$$

where $\delta = \varepsilon_0/2$ and $k_0(n) = \max_{i=1, \dots, m-1} (u_{i+1} - u_i)$.

Proof. Let $u_i \in \mathcal{U}_n$, and let $\mathbf{Z}_n^{(u_i)}$ be a random vector with law $\mathbb{P}^{(u_i)}$. By (3.8),

$$\ell_n(u_{i+1}) = \mathbb{E}\left(L_n\left(\tilde{\mathbf{Z}}_n^{(u_i)}\right)\right).$$

Hence,

$$\ell_n(u_{i+1}) - \ell_n(u_i) = \mathbb{E}\left(L_n\left(\tilde{\mathbf{Z}}_n^{(u_i)}\right)\right) - \mathbb{E}\left(L_n\left(\mathbf{Z}_n^{(u_i)}\right)\right) = \mathbb{E}\left[\mathbb{E}\left(L_n\left(\tilde{\mathbf{Z}}_n^{(u_i)}\right) - L_n\left(\mathbf{Z}_n^{(u_i)}\right) \mid \mathbf{Z}_n^{(u_i)}\right)\right].$$

Let ε_0 be as in (ii), and let $B_n(\varepsilon_0) \subset \mathcal{X}_n \times \mathcal{X}_n$ be the set of outcomes of \mathbf{Z}_n such that

$$\left\{\mathbb{E}\left(L_n\left(\tilde{\mathbf{Z}}_n\right) - L_n\left(\mathbf{Z}_n\right) \mid \mathbf{Z}_n\right) \geq \varepsilon_0\right\} = \{\mathbf{Z}_n \in B_n(\varepsilon_0)\}. \quad (3.13)$$

By (3.9), for any pair of sequences z_n , when applying the transformation \mathcal{R} , the score can decrease by at most $-A$. Hence

$$\mathbb{E}\left[\mathbb{E}\left(L_n\left(\tilde{\mathbf{Z}}_n^{(u_i)}\right) - L_n\left(\mathbf{Z}_n^{(u_i)}\right) \mid \mathbf{Z}_n^{(u_i)}\right)\right] \geq \varepsilon_0 \mathbb{P}\left(\mathbf{Z}_n^{(u_i)} \in B_n(\varepsilon_0)\right) - A \mathbb{P}\left(\mathbf{Z}_n^{(u_i)} \notin B_n(\varepsilon_0)\right).$$

By definition, $\mathbb{P}(\mathbf{Z}_n \notin B_n(\varepsilon_0)) = \Delta_n(\varepsilon_0)$. Therefore, by (3.10) and (3.11),

$$\mathbb{P}\left(\mathbf{Z}_n^{(u_i)} \notin B_n(\varepsilon_0)\right) = \mathbb{P}\left(\mathbf{Z}_n \notin B_n(\varepsilon_0) \mid U_n = u_i\right) \leq \frac{\Delta_n(\varepsilon_0)}{\mathbb{P}(U_n = u_i)} \leq \frac{\Delta_n(\varepsilon_0)}{\varphi(n)} = o(1), \quad n \rightarrow \infty.$$

Choose n_0 large enough (depending on ε_0 and A) so that for any $n > n_0$,

$$\varepsilon_0 \left(1 - \frac{\Delta_n(\varepsilon_0)}{\varphi(n)}\right) - A \frac{\Delta_n(\varepsilon_0)}{\varphi(n)} \geq \frac{\varepsilon_0}{2}.$$

Therefore, for any $n > n_0$ and any $u_i \in \mathcal{U}_n$,

$$\ell_n(u_{i+1}) - \ell_n(u_i) \geq \frac{\varepsilon_0}{2} =: \delta.$$

This proves (3.2), which entails (3.4) and thus (3.12). The proof is now complete. \square

Remark 3.5. The assumption (3.9) typically holds for any meaningful function \mathbf{u} . The difficulties in applying Theorem 3.4 lie in finding \mathbf{u} , \mathcal{R} and \mathcal{U}_n , such that (3.10) holds for a given ε_0 . Since typically for every $u \in \mathcal{S}_n$, $\mathbb{P}(U_n = u) \rightarrow 0$, $\Delta_n(\varepsilon_0)$ has to converge to zero sufficiently fast as $n \rightarrow \infty$. The larger the probability $\mathbb{P}(U_n \in \mathcal{U}_n)$ is (recall that it has to be bounded away from zero), the smaller $\varphi(n)$ is, and the faster the convergence to zero of $\Delta_n(\varepsilon_0)$ is.

3.3 Uniform Approximation

As mentioned in Remark 3.5, the most important assumption in Theorem 3.4 is the existence of the random transformation \mathcal{R} and of the corresponding random variable U_n such that (for some $\varepsilon_0 > 0$) the probability $\Delta_n(\varepsilon_0)$ converges to zero sufficiently fast. For all purposes, this is often a very specific requirement and so far, the existence of suitable \mathcal{R} and U_n has only been shown for very specific models (cf. [17], [19], [21] and [16]). It is therefore important to relax the assumption of fast convergence, which is the goal of the present subsection. Below, the condition (3.10) is replaced by an arbitrarily slowly convergent sequence $\Delta_n(\varepsilon_0) \rightarrow 0$.

Theorem 3.6. *For every $n \in \mathbb{N}$, let there exist a random variable $U_n = \mathbf{u}(\mathbf{Z}_n)$, a set $\mathcal{U}_n = \{u_1, \dots, u_{m(n)}\} \subset \mathcal{S}_n$, a random transformation \mathcal{R} , a positive function φ on \mathbb{N} , and a constant $A > 0$, such that (3.8), (3.9) and (3.11) are satisfied. Moreover, let there exist a universal constant $c > 0$, such that*

$$m(n) \geq c \varphi^{-1}(n), \quad \text{for all } n \in \mathbb{N}. \quad (3.14)$$

If for an $\varepsilon_0 > 0$, $\Delta_n(\varepsilon_0) \rightarrow 0$, then for n large enough,

$$\mathbb{E}(\Phi(|L_n(\mathbf{Z}_n) - \mu_n|)) \geq \sum_{j=1}^{\frac{\varepsilon_0}{8} \varphi^{-1}(n)} \Phi\left(\frac{\varepsilon_0}{2} \left(\frac{c}{8\varphi(n)} + j\right)\right) \varphi(n) + \frac{c}{8} \Phi\left(\frac{\varepsilon_0 c}{16\varphi(n)}\right) \geq \frac{c}{4} \Phi\left(\frac{\varepsilon_0 c}{16\varphi(n)}\right).$$

Proof. Let ε_0 be as in the assumption of the theorem and let the set $\mathcal{U}_n^0 \subseteq \mathcal{S}_n$ be defined as follows:

$$u \in \mathcal{U}_n^0 \Leftrightarrow \mathbb{P}(\mathbf{Z}_n \notin B_n(\varepsilon_0) | U_n = u) \leq \sqrt{\Delta_n(\varepsilon_0)}, \quad (3.15)$$

where $B_n(\varepsilon_0)$ is given by (3.13). Thus, by the very definition of $\Delta_n(\varepsilon_0)$,

$$\Delta_n(\varepsilon_0) = \mathbb{P}(\mathbf{Z}_n \notin B_n(\varepsilon_0)) \geq \sum_{u \notin \mathcal{U}_n^0} \mathbb{P}(\mathbf{Z}_n \notin B_n(\varepsilon_0) | U_n = u) \mathbb{P}(U_n = u) \geq \sqrt{\Delta_n(\varepsilon_0)} \mathbb{P}(U_n \notin \mathcal{U}_n^0),$$

which implies that

$$\mathbb{P}(U_n \notin \mathcal{U}_n^0) \leq \sqrt{\Delta_n(\varepsilon_0)}. \quad (3.16)$$

When $u_i \in \mathcal{U}_n \cap \mathcal{U}_n^0$, using an argument as in the proof of Theorem 3.4, and by (3.9) and (3.15),

$$\ell_n(u_{i+1}) - \ell_n(u_i) \geq \varepsilon_0 \left(1 - \sqrt{\Delta_n(\varepsilon_0)}\right) - A\sqrt{\Delta_n(\varepsilon_0)} \geq \frac{\varepsilon_0}{2}, \quad (3.17)$$

provided that $n > n_1$, for some positive integer n_1 large enough. When $u_i \notin \mathcal{U}_n \cap \mathcal{U}_n^0$, by (3.9),

$$\ell_n(u_{i+1}) - \ell_n(u_i) \geq -A. \quad (3.18)$$

Assuming $n > n_1$, we are interested in the set $\mathcal{U}_n \cap \mathcal{U}_n^0$ which can be represented as the union of disjoint subintervals of $\mathcal{U}_n = \{u_1, \dots, u_m\}$, i.e., $\mathcal{U}_n \cap \mathcal{U}_n^0 = \cup_{j=1}^N I_j$, where for $j = 1, \dots, N$,

$$I_j = \left\{u_1^{(j)}, \dots, u_{r_j}^{(j)}\right\}, \quad \text{with } u_1^{(1)} < \dots < u_{r_1}^{(1)} < u_1^{(2)} < \dots < u_{r_2}^{(2)} < \dots < u_1^{(N)} < \dots < u_{r_N}^{(N)},$$

are disjoint subintervals of \mathcal{U}_n . Clearly, the number N of such intervals as well as the intervals I_1, \dots, I_N themselves depend on n . On each interval I_j , (3.17) implies that the function ℓ_n increases with a slope of at least $\varepsilon_0/2$, and moreover, note that if $u_i = u_{r_j}^{(j)}$, then $\ell_n(u_{i+1}) \notin I_j$. Note also that the approach in proving Theorem 3.4 largely applied because \mathcal{U}_n was a single interval where ℓ_n increases. In the present, case the set $\mathcal{U}_n \cap \mathcal{U}_n^0$ consists of several intervals and between them the function ℓ_n does not necessarily increase. At first, it might appear that the approach of the previous subsection would work, when replacing

$$\sum_{u_i \in \mathcal{U}_n} \Phi(\delta |u_i - a_n|) \mathbb{P}(U_n = u_i)$$

with

$$\sum_{u_i \in \mathcal{U}_n \cap \mathcal{U}_n^0} \Phi(\delta |u_i - a_n|) \mathbb{P}(U_n = u_i) = \sum_{j=1}^N \sum_{u_i \in I_j} \Phi(\delta |u_i - a_n|) \mathbb{P}(U_n = u_i).$$

Unfortunately, a_n would now depend on I_j , and we would have to deal with the sums

$$\sum_{j=1}^N \sum_{u_i \in I_j} \Phi\left(\delta \left|u_i - a_n^{(j)}\right|\right) \mathbb{P}(U_n = u_i).$$

To bypass this problem, we proceed differently. Consider the sets

$$J_j := \left\{ \ell_n(u_1^{(j)}), \dots, \ell_n(u_{r_j}^{(j)}) \right\}, \quad j = 1, \dots, N,$$

i.e., J_j is the image of I_j under ℓ_n . By (3.17), all elements of J_j are at least $\varepsilon_0/2$ -apart from each other (the intervals J_j might overlap, although the intervals I_j do not). By (3.18),

$$\sum_{u_i \in \mathcal{U}_n \setminus \mathcal{U}_n^0} (\ell_n(u_{i+1}) - \ell_n(u_i)) \geq -A \cdot \text{Card}(\mathcal{U}_n \setminus \mathcal{U}_n^0) =: -A m_0(n),$$

implying that the sum of the lengths of the (integer) intervals J_j differs from the length of $J := \bigcup_{j=1}^N J_j$ by at most $A m_0(n)$. Formally,

$$\sum_{j=1}^N \left(\ell_n(u_{r_j}^{(j)}) - \ell_n(u_1^{(j)}) \right) - |J| \leq A m_0(n),$$

where $|J|$ denotes the length of J , i.e., the difference between the largest element and the smallest element of J , and $\ell_n(u_{r_j}^{(j)}) - \ell_n(u_1^{(j)})$ is the length of each J_j . The number of elements $\varepsilon_0/2$ -apart needed for covering a (real) interval with length $A m_0(n)$ is at most $2A m_0(n)/\varepsilon_0 + 1$. Hence, at most $2A m_0(n)/\varepsilon_0 + 1$ elements $\varepsilon_0/2$ -apart will be lost due to overlappings, which implies that the set J contains at least

$$|\mathcal{U}_n| - \frac{2A m_0(n)}{\varepsilon_0} - 1 = m(n) - \frac{2A m_0(n)}{\varepsilon_0} - 1$$

elements which are (at least) $\varepsilon_0/2$ -apart from each other. Using (3.11) and (3.16),

$$\sqrt{\Delta_n(\varepsilon_0)} \geq \mathbb{P}(U_n \in \mathcal{U}_n \setminus \mathcal{U}_n^0) = \sum_{u \in \mathcal{U}_n \setminus \mathcal{U}_n^0} \mathbb{P}(U_n = u) \geq m_0(n) \varphi(n),$$

and therefore,

$$m_0(n) \leq \sqrt{\Delta_n(\varepsilon_0)} \varphi^{-1}(n).$$

Recalling the assumption (3.14), there exists $n_2 \in \mathbb{N}$ such that for all $n > n_2$,

$$m(n) - \frac{2A m_0(n)}{\varepsilon_0} - 1 \geq m(n) - \frac{2A \sqrt{\Delta_n(\varepsilon_0)}}{\varepsilon_0 \varphi(n)} \geq \frac{c\varepsilon_0 - 2A \sqrt{\Delta_n(\varepsilon_0)}}{\varepsilon_0 \varphi(n)} \geq \frac{c}{2} \varphi^{-1}(n). \quad (3.19)$$

Now let

$$\mathcal{B}_n := \left\{ u \in \mathcal{U}_n : |\ell_n(u) - \mu_n| \geq \frac{\varepsilon_0 c}{16\varphi(n)} \right\} = \mathcal{B}_n^+ \cup \mathcal{B}_n^-,$$

where

$$\mathcal{B}_n^+ := \left\{ u \in \mathcal{U}_n : \ell_n(u) - \mu_n \geq \frac{\varepsilon_0 c}{16\varphi(n)} \right\} \quad \text{and} \quad \mathcal{B}_n^- := \left\{ u \in \mathcal{U}_n : \ell_n(u) - \mu_n \leq -\frac{\varepsilon_0 c}{16\varphi(n)} \right\}.$$

Since the interval

$$\left(\mu_n - \frac{\varepsilon_0 c}{16\varphi(n)}, \mu_n + \frac{\varepsilon_0 c}{16\varphi(n)} \right)$$

contains at most $c\varphi^{-1}(n)/4$ elements which are $\varepsilon_0/2$ -apart from each other, and since the set J contains at least $c\varphi^{-1}(n)/2$ elements which are $\varepsilon_0/2$ -apart from each other, it follows that \mathcal{B}_n contains at least $c\varphi^{-1}(n)/4$ elements (the values of ℓ_n being $\varepsilon_0/2$ -apart from each other). This implies that at least one of the two sets \mathcal{B}_n^+ and \mathcal{B}_n^- , say \mathcal{B}_n^+ , contains at least $c\varphi^{-1}(n)/8$ elements whose ℓ_n -images are at least $\varepsilon_0/2$ -apart from each other. Let $\tilde{\mathcal{B}}_n$ be the union of those $c\varphi^{-1}(n)/8$ elements in \mathcal{B}_n^+ . Then the set $\mathcal{B}_n \setminus \tilde{\mathcal{B}}_n$ consists of at least $c\varphi^{-1}(n)/8$ elements (whose ℓ_n -images are at least $\varepsilon_0/2$ -apart from each other). Hence, by the monotonicity of Φ ,

$$\begin{aligned} \sum_{i=1}^{m(n)} \Phi(|\ell_n(u_i) - \mu_n|) \mathbb{P}(U_n = u_i) &\geq \sum_{i=1}^{m(n)} \Phi(|\ell_n(u_i) - \mu_n|) \varphi(n) \geq \sum_{u \in \mathcal{B}_n} \Phi(|\ell_n(u) - \mu_n|) \varphi(n) \\ &\geq \sum_{u \in \tilde{\mathcal{B}}_n} \Phi(|\ell_n(u) - \mu_n|) \varphi(n) + \sum_{u \in \mathcal{B}_n \setminus \tilde{\mathcal{B}}_n} \Phi(|\ell_n(u) - \mu_n|) \varphi(n) \\ &\geq \sum_{j=1}^{\frac{c}{8}\varphi^{-1}(n)} \Phi\left(\frac{\varepsilon_0}{2} \left(\frac{c}{8\varphi(n)} + j\right)\right) \varphi(n) + \Phi\left(\frac{\varepsilon_0 c}{16\varphi(n)}\right) \frac{c}{8} \varphi^{-1}(n) \varphi(n) \\ &\geq \sum_{j=1}^{\frac{c}{8}\varphi^{-1}(n)} \Phi\left(\frac{\varepsilon_0}{2} \left(\frac{c}{8\varphi(n)} + j\right)\right) \varphi(n) + \Phi\left(\frac{\varepsilon_0 c}{16\varphi(n)}\right) \frac{c}{8} \\ &\geq \frac{c}{4} \Phi\left(\frac{\varepsilon_0 c}{16\varphi(n)}\right). \end{aligned}$$

The inequality (3.5) now finishes the proof. \square

Remark 3.7. In both Theorem 3.4 and Theorem 3.6, the thresholds, on large n , are independent of the choice of the convex function Φ . In Theorem 3.4, the threshold n_0 depends on the sequence $\Delta_n(\varepsilon_0)/\varphi(n)$, but not on Φ . In Theorem 3.6, the threshold on n is obtained in two steps: the first threshold n_1 is to ensure the validity of (3.17), while the second threshold n_2 is to guarantee the validity of the last inequality in (3.19). Both steps are independent of the choice of Φ .

Remark 3.8. The assumption $m(n) = |\mathcal{U}_n| \geq c\varphi^{-1}(n)$ (for all $n \in \mathbb{N}$), together with (3.11), implies that for every $u \in \mathcal{U}_n$, $\mathbb{P}(U_n = u) \geq \varphi(n) \geq cm^{-1}(n)$. Hence, \mathcal{U}_n cannot be too large, since all the atoms in \mathcal{U}_n have roughly equal masses, justifying our naming this approach “uniform approximation”. On the other hand, these two assumptions imply that $\mathbb{P}(U \in \mathcal{U}_n) \geq c$, so that \mathcal{U}_n cannot be too small either. In the next section, we will see that the set \mathcal{U}_n satisfying the requirements of Theorem 3.6 does exist, but its choice must be made more carefully when compared to the one in Theorem 3.4. This is the price to pay for the arbitrarily slow convergence of $\Delta_n(\varepsilon_0)$.

4 The Binary LCS Case

As an applications of the general results of the previous section, in the rest of the paper, we consider the binary alphabet $\mathcal{A} = \{0, 1\}$, i.e., \mathbf{X}_n and \mathbf{Y}_n are two independent i.i.d. sequences such that $\mathbb{P}(X_1 = 1) = \mathbb{P}(Y_1 = 1) = p$, $p \in (0, 1)$, and $L_n(\mathbf{X}_n; \mathbf{Y}_n)$ is the length of the LCS of \mathbf{X}_n and \mathbf{Y}_n . Let $u(\mathbf{z}_n)$ count the zeros in $\mathbf{z}_n = (\mathbf{x}_n; \mathbf{y}_n)$. Thus U_n is equal to the number of zeros in $\mathbf{Z}_n = (\mathbf{X}_n; \mathbf{Y}_n)$ and therefore $U_n \sim B(2n, q)$, where $q = 1 - p$, and its support is $\mathcal{S}_n = \{0, 1, \dots, 2n\}$.

Let the random transformation \mathcal{R} be defined as follows: given \mathbf{z}_n , choose randomly (with uniform probability) a one and turn it into a zero. With this transformation, the condition (3.8) holds for any $u = 0, \dots, 2n - 1$. To see this, for any $u = 0, \dots, 2n$, let $\mathcal{A}_n(u) \subseteq \{0, 1\}^{2n}$ consist of all the binary sequences containing exactly u zeros. For any $\mathbf{z}_n \in \mathcal{A}_n(u)$,

$$\mathbb{P}(\mathbf{Z}_n = \mathbf{z}_n | U_n = u) = \frac{\mathbb{P}(\mathbf{Z}_n = \mathbf{z}_n, U_n = u)}{\mathbb{P}(U_n = u)} = \frac{\mathbb{P}(\mathbf{Z}_n = \mathbf{z}_n)}{\mathbb{P}(U_n = u)} = \frac{q^u(1-q)^{2n-u}}{\binom{u}{2n} q^u(1-q)^{2n-u}} = \binom{u}{2n}^{-1}.$$

In other words, $\mathbb{P}_{(u)}$ is the uniform distribution on $\mathcal{A}_n(u)$. Now, for any such u , let $\mathbf{Z}_n^{(u)}$ be a random vector with law $\mathbb{P}^{(u)}$, then $\tilde{\mathbf{Z}}_n^{(u)} = \mathcal{R}(\mathbf{Z}_n^{(u)})$ is supported on $\mathcal{A}_n(u+1)$. For any $\mathbf{z}_n \in \mathcal{A}_n(u+1)$, let $0 \leq i_1 < \dots < i_{u+1} \leq 2n$ be the positions of the zeros in \mathbf{z}_n , and let $\hat{\mathbf{z}}_n^{(i_j)}$, $j = 1, \dots, u+1$, be the sequence in $\mathcal{A}_n(u)$ obtained by changing the zero of \mathbf{z}_n at the position i_j to a one. Then,

$$\mathbb{P}(\tilde{\mathbf{Z}}_n^{(u)} = \mathbf{z}_n) = \sum_{j=1}^{u+1} \mathbb{P}(\tilde{\mathbf{Z}}_n^{(u)} = \mathbf{z}_n | \mathbf{Z}_n^{(u)} = \hat{\mathbf{z}}_n^{(i_j)}) \mathbb{P}(\mathbf{Z}_n^{(u)} = \hat{\mathbf{z}}_n^{(i_j)}) = \frac{u+1}{2n-u} \binom{u}{2n}^{-1} = \binom{u+1}{2n}^{-1} = \mathbb{P}^{(u+1)}(\mathbf{z}_n).$$

Finally note that since we are considering the LCS, any single-entry change in \mathbf{z}_n changes the score by at most one. Hence, $A = 1$ in (3.9) and

$$-1 \leq \ell_n(u+1) - \ell_n(u) \leq 1, \quad \text{for any } u \in \{0, 1, \dots, 2n-1\}. \quad (4.1)$$

4.1 Lower Bounds Under Fast Convergence

The following theorem is proved in [19, Theorem 2.2] (see also [16, Theorem 2.1] for a quantitative version with an arbitrary finite alphabet size).

Theorem 4.1. *There exist positive constants ε_1 and ε_2 with $\varepsilon_1 > \varepsilon_2$, and a set $B_n \subseteq \mathcal{A}^n \times \mathcal{A}^n$ such that for every $\mathbf{z}_n \in B_n$,*

$$\mathbb{P}\left(L_n(\tilde{\mathbf{Z}}_n) - L_n(\mathbf{Z}_n) = 1 \mid \mathbf{Z}_n = \mathbf{z}_n\right) \geq \varepsilon_1, \quad \mathbb{P}\left(L_n(\tilde{\mathbf{Z}}_n) - L_n(\mathbf{Z}_n) = -1 \mid \mathbf{Z}_n = \mathbf{z}_n\right) \leq \varepsilon_2.$$

Moreover, there exists $p_0 > 0$, such that for every $0 < p < p_0$,

$$\mathbb{P}(\mathbf{Z}_n \in B_n) \geq 1 - e^{-c_1 n},$$

where $c_1 > 0$ does not depend on n , but may depend on p .

Hence (3.7) holds with $\varepsilon_0 = \varepsilon_1 - \varepsilon_2$ and $\Delta_n(\varepsilon_0) \leq e^{-c_1 n}$, for n large enough. Furthermore, as shown above, (3.8) also holds for $u_i = 0, 1, \dots, 2n - 1$.

In what follows, we will apply the fast convergence result (Theorem 3.4) to two possible choices of \mathcal{U}_n , and then compare the corresponding lower bounds on moments and exponential moments of $|L_n(\mathbf{Z}_n) - \mu_n|$.

4.1.1 The Standard \mathcal{U}_n

Often (cf. [19], [21] and [16]) the following choice of \mathcal{U}_n is used:

$$\mathcal{U}_n = [2nq - \sqrt{2n}, 2nq + \sqrt{2n}] \cap \mathcal{S}_n.$$

This “standard \mathcal{U}_n ” is such that $k_0 = 1$ (recall (3.3)). By the de Moivre-Laplace local limit theorem (cf. [27, pp. 56]), there exists a universal constant $b = b(p) > 0$, which is independent of n , but depends on p , such that for n large enough,

$$\mathbb{P}(U_n = u) \geq \frac{1}{b\sqrt{n}}, \quad \text{for any } u \in \mathcal{U}_n. \quad (4.2)$$

The (best) constant b depends on the choice of \mathcal{U}_n , and it is easy to see (see (4.10) below) that for the standard \mathcal{U}_n as above, $b \geq e^2 \sqrt{\pi}$. Without loss of generality, assume that $2nq$ is a positive integer (if not, replace $2nq$ in the definition of \mathcal{U}_n with $\lceil 2nq \rceil$). Hence, with $\varphi(n) = 1/b\sqrt{n}$, the condition (3.10) is verified and thus, all the assumptions of Theorem 3.4 are satisfied. Therefore, recalling that $\delta = \varepsilon_0/2$, $\mathbb{E}(\Phi(|L_n(\mathbf{Z}_n) - \mu_n|))$ is lower-bounded by

$$\begin{aligned} \sum_{u_i \in \mathcal{U}_n} \Phi(\delta |u_i - a_n|) \mathbb{P}(U_n = u_i) &\geq \sum_{u_i \in \mathcal{U}_n} \Phi(\delta |u_i - a_n|) \frac{1}{b\sqrt{n}} = \frac{1}{b\sqrt{n}} \sum_{k=2qn-\sqrt{2n}}^{2qn+\sqrt{2n}} \Phi(\delta |k - a_n|) \\ &\geq \frac{1}{b\sqrt{n}} \sum_{j=-\sqrt{2n}}^{\sqrt{2n}} \Phi(\delta |j|) = \frac{1}{b\sqrt{n}} \left(\Phi(0) + 2 \sum_{j=1}^{\sqrt{2n}} \Phi\left(\frac{\varepsilon_0 j}{2}\right) \right), \end{aligned} \quad (4.3)$$

where the last inequality is obtained by minimizing over a_n and since $2nq$ is a positive integer.

Example 4.2. For $\Phi(x) = |x|^r$, $r \geq 1$, then (4.3) becomes

$$\frac{\varepsilon_0^r}{2^{r-1}b\sqrt{n}} \sum_{j=1}^{\sqrt{2n}} j^r \geq \frac{\varepsilon_0^r}{2^{r-1}b\sqrt{n}} \int_0^{\sqrt{2n}} x^r dx = \frac{\varepsilon_0^r (2n)^{(r+1)/2}}{(r+1)2^{r-1}b\sqrt{n}} = \frac{2^{(3-r)/2} \varepsilon_0^r}{b(r+1)} n^{r/2}.$$

Hence, for n large enough (independent of r , see Remark 3.7),

$$\mathbb{E}(|L_n(\mathbf{Z}_n) - \mu_n|^r) \geq d_1(r) n^{r/2}, \quad \text{where } d_1(r) = d_1(r, p, \varepsilon_0) := \frac{2^{(3-r)/2} \varepsilon_0^r}{b(r+1)}. \quad (4.4)$$

Example 4.3. For $\Phi(x) = e^{tx}$, $t > 0$, then the second summation in (4.3) becomes

$$\sum_{j=1}^{\sqrt{2n}} \Phi\left(\frac{\varepsilon_0 j}{2}\right) = \sum_{j=1}^{\sqrt{2n}} e^{t\varepsilon_0 j/2} = \sum_{j=1}^{\sqrt{2n}} \rho_t^j = \frac{\rho_t - \rho_t^{\sqrt{2n}+1}}{1 - \rho_t}, \quad \text{where } \rho_t := e^{\varepsilon_0 t/2}.$$

Hence, for n large enough,

$$\mathbb{E}\left(e^{|L_n(\mathbf{Z}_n) - \mu_n|t}\right) \geq \frac{1}{b\sqrt{n}} \left(1 + \frac{2\rho_t - 2\rho_t^{\sqrt{2n}+1}}{1 - \rho_t}\right) = \frac{\rho_t(2\rho_t^{\sqrt{2n}} - \rho_t^{-1} - 1)}{b(\rho_t - 1)\sqrt{n}} = \frac{2\rho_t^{\sqrt{2n}} - 1 - \rho_t^{-1}}{b(1 - \rho_t^{-1})\sqrt{n}}. \quad (4.5)$$

Now, for any $t > 0$, there exists n large enough (depending on t) such that

$$\mathbb{E}\left(e^{|L_n(\mathbf{Z}_n) - \mu_n|t}\right) \geq \frac{2\rho_t}{b(\rho_t - 1)} \rho_t^{\sqrt{n}}.$$

Taking $t = s/\sqrt{n}$, with $s > 0$, gives a lower bound on the moment generating function of $|V_n| = |L_n(\mathbf{Z}_n) - \mu_n|/\sqrt{n}$. Moreover, by (4.5),

$$\liminf_{n \rightarrow \infty} \mathbb{E}\left(e^{s|V_n|}\right) \geq \liminf_{n \rightarrow \infty} \frac{2\rho_t^{\sqrt{2n}} - 1 - \rho_t^{-1}}{b(1 - \rho_t^{-1})\sqrt{n}} = \frac{4}{b\varepsilon_0 s} \left(e^{\varepsilon_0 s/\sqrt{2}} - 1\right). \quad (4.6)$$

Since

$$\lim_{s \downarrow 0} \frac{1}{\varepsilon_0 s} \left(e^{\varepsilon_0 s/\sqrt{2}} - 1\right) = \frac{1}{\sqrt{2}},$$

the right-hand side of (4.6) approaches $2\sqrt{2}/b$ as $s \downarrow 0$. Since $b \geq \sqrt{\pi}e^2$, then $2\sqrt{2} < b$, showing the deficiency of the obtained bound, since we naturally expect the right hand side of (4.6) to approach one as $s \downarrow 0$.

Remark 4.4. Since $|V_n|$ is non-negative, a series expansion, together with lower bounds on moments of $|V_n|$, can also lower-bound its moment generating function. Indeed, by (4.4), for n large enough (independent of r),

$$\mathbb{E}\left(e^{s|V_n|}\right) = 1 + \sum_{r=1}^{\infty} \mathbb{E}(|V_n|^r) \frac{s^r}{r!} \geq 1 + \frac{2\sqrt{2}}{b} \sum_{r=1}^{\infty} \frac{\left(\frac{\varepsilon_0 s}{\sqrt{2}}\right)^r}{(r+1)!} = 1 - \frac{2\sqrt{2}}{b} + \frac{4}{b\varepsilon_0 s} \left(e^{\varepsilon_0 s/\sqrt{2}} - 1\right). \quad (4.7)$$

This gives a lower bound similar to (4.6), except for the additive constant $1 - 2\sqrt{2}/b$, which ensures that the right-hand side of (4.7) tends to one as $s \downarrow 0$.

Note finally that (4.6) is computed using the lower bound in (4.3) with $\Phi(x) = e^{tx}$ and $t = s/\sqrt{n}$:

$$\frac{1}{b\sqrt{n}} \sum_{j=-\sqrt{2n}}^{\sqrt{2n}} e^{t\varepsilon_0 |j|/2} = \frac{1}{b\sqrt{n}} \sum_{j=-\sqrt{2n}}^{\sqrt{2n}} \sum_{r=0}^{\infty} \frac{1}{r!} \left(\frac{t\varepsilon_0 |j|}{2}\right)^r = \frac{2\sqrt{2n} + 1}{b\sqrt{n}} + \sum_{r=1}^{\infty} \left[\frac{\varepsilon_0^r}{2^{r-1}b\sqrt{n}r!} \left(\frac{s}{\sqrt{n}}\right)^r \sum_{j=1}^{\sqrt{2n}} j^r \right].$$

Above, the constant term (i.e., $r = 0$) converges to $2\sqrt{2}/b$ as $n \rightarrow \infty$, which is different from the constant term (equal to 1) in the series expansion in (4.7). The lower bound on the last series above can be computed using (4.4):

$$\sum_{r=1}^{\infty} \left[\frac{\varepsilon_0^r}{2^{r-1}b\sqrt{n}r!} \left(\frac{s}{\sqrt{n}}\right)^r \sum_{j=1}^{\sqrt{2n}} j^r \right] \geq \sum_{r=1}^{\infty} \frac{1}{r!} \left(\frac{s}{\sqrt{n}}\right)^r \frac{2^{(3-r)/2} \varepsilon_0^r}{b(r+1)} n^{\frac{r}{2}} = -\frac{2\sqrt{2}}{b} + \frac{4}{b\varepsilon_0 s} \left(e^{\varepsilon_0 s/\sqrt{2}} - 1\right).$$

Therefore, the difference in the constants in (4.6) and (4.7) stems from the different constant terms ($r = 0$) in the respective series expansions.

4.1.2 The Extended \mathcal{U}_n

The bounds presented in the previous subsection give the correct order for the centered absolute moments (given that p is small enough), i.e., $\mathbb{E}(|L_n - \mu_n|^r) \asymp O(n^{r/2})$. On the other hand, as argued at the end of Example 4.3, the lower bound on the moment generating function of $|V_n|$ can be improved. A way to do so (also valid for the moments of $|V_n|$) is to extend the set \mathcal{U}_n and to refine the lower bound (4.2). In the present subsection, we show how the refined lower bound yields better approximations.

In what follows, take $\beta \in (1/2, 2/3)$, and redefine \mathcal{U}_n as follows

$$\mathcal{U}_n = \left[2nq - (2n)^\beta, 2nq + (2n)^\beta\right] \cap \mathcal{S}_n = \left\{k \in \mathbb{Z} : |k - 2nq| \leq (2n)^\beta\right\}. \quad (4.8)$$

In the sequel, this \mathcal{U}_n is called the “extended” \mathcal{U}_n , and note that again $k_0 = 1$ (recall (3.3)). Since $(2n)^\beta = o(n^{2/3})$, it follows from the local de Moivre-Laplace theorem (cf. [27, pp. 56]) that

$$\sup_{k: |k-2nq| \leq (2n)^\beta} \left| \frac{\mathbb{P}(U_n = k)}{\phi_n(k)} - 1 \right| \rightarrow 0,$$

where

$$\phi_n(k) := \frac{1}{2\sqrt{\pi pqn}} \exp\left(-\frac{(k - 2nq)^2}{4npq}\right), \quad \text{and } q = 1 - p.$$

Hence, for every $\epsilon > 0$, there exists $N \in \mathbb{N}$ such that, for any $n > N$,

$$(1 + \epsilon) \phi_n(k) \geq \mathbb{P}(U_n = k) \geq (1 - \epsilon) \phi_n(k), \quad \text{for all } k \in \mathbb{Z} \text{ with } |k - 2nq| \leq (2n)^\beta. \quad (4.9)$$

In particular,

$$\mathbb{P}(U_n = k) \geq \frac{(1 - \epsilon)}{2\sqrt{\pi pqn}} \exp\left(-\frac{(2n)^{2\beta-1}}{2pq}\right) =: \varphi(n). \quad (4.10)$$

Since $2\beta - 1 < 1$, it follows that

$$\varphi^{-1}(n) e^{-c_1 n} \rightarrow 0, \quad \text{whenever } c_1 > 0.$$

On the other hand, by Theorem 4.1, there exists $\varepsilon_0 > 0$ and $c_1 > 0$, such that $\Delta_n(\varepsilon_0) \leq e^{-c_1 n}$, for any $n \in \mathbb{N}$ large enough. Thus, the condition (3.10) holds and therefore, all the assumptions of Theorem 3.4 are verified.

Thus, recalling that $\delta = \varepsilon_0/2$ and $a_n \in [2nq - (2n)^\beta, 2nq + (2n)^\beta]$,

$$\mathbb{E}(\Phi(|L_n(\mathbf{Z}_n) - \mu_n|)) \geq \sum_{k \in \mathcal{U}_n} \Phi(\delta |k - a_n|) \mathbb{P}(U_n = k) \geq (1 - \epsilon) \sum_{k: |k-2nq| \leq (2n)^\beta} \Phi(\delta |k - a_n|) \phi_n(k),$$

for n large enough, where the inequality follows from (4.9). From the symmetry of $\Phi(\delta|\cdot|)$, the right-hand side above is minimized at $a_n = 2nq$. Moreover, let

$$j_k(n) := \frac{(k - 2nq)}{\sqrt{2n}}, \quad \text{for } k \in \mathbb{Z} \text{ with } |k - 2nq| \leq (2n)^\beta.$$

Clearly, $(j_k(n))_{k \in \mathcal{U}_n}$ forms a $1/\sqrt{2n}$ -net over the interval $[-(2n)^{\beta-1/2}, (2n)^{\beta-1/2}]$. Hence,

$$\begin{aligned} \mathbb{E}(\Phi(|L_n(\mathbf{Z}_n) - \mu_n|)) &\geq (1 - \epsilon) \sum_{k: |k-2nq| \leq (2n)^\beta} \Phi(\delta |k - 2nq|) \phi_n(k) \\ &= (1 - \epsilon) \sum_{k: |k-2nq| \leq (2n)^\beta} \frac{1}{2\sqrt{\pi pq n}} \Phi(\delta \sqrt{2n} |j_k(n)|) \exp\left(-\frac{(j_k(n))^2}{2pq}\right). \end{aligned} \quad (4.11)$$

Example 4.5. For $\Phi(x) = |x|^r$, $r \geq 1$, since $n^{\beta-1/2} \rightarrow \infty$, as $n \rightarrow \infty$,

$$\lim_{n \rightarrow \infty} \frac{1}{\sqrt{2\pi pq}} \sum_{k: |k-2nq| \leq (2n)^\beta} \frac{1}{\sqrt{2n}} |j_k(n)|^r \exp\left(-\frac{(j_k(n))^2}{2pq}\right) = \frac{1}{\sqrt{2\pi pq}} \int_{-\infty}^{\infty} |x|^r e^{-x^2/(2pq)} dx.$$

Since $\epsilon > 0$ is arbitrary, (4.11) leads to

$$\liminf_{n \rightarrow \infty} \mathbb{E}(|V_n|^r) = \liminf_{n \rightarrow \infty} \frac{\mathbb{E}(|L_n(\mathbf{Z}_n) - \mu_n|^r)}{n^{r/2}} \geq 2^{r/2} \delta^r \mathbb{E}(|\xi|^r) = \frac{\varepsilon_0^r}{\sqrt{\pi}} (pq)^{r/2} \Gamma\left(\frac{r+1}{2}\right) =: d_2(r),$$

where $\xi \sim \mathcal{N}(0, pq)$. Therefore, for $n \in \mathbb{N}$ large enough,

$$\mathbb{E}(|L_n(\mathbf{Z}_n) - \mu_n|^r) \geq \frac{\varepsilon_0^r}{\sqrt{\pi}} (pq)^{r/2} \Gamma\left(\frac{r+1}{2}\right) n^{r/2}.$$

Let us compare this last bound with (4.4) of Example 4.2. From (4.9),

$$\frac{1}{b} \leq \frac{1}{2\sqrt{\pi pq}} e^{-1/(2pq)} \quad \text{implies that} \quad d_1(r) \leq \frac{2^{\frac{1-r}{2}} \varepsilon_0^r}{(r+1)\sqrt{\pi pq}} e^{-1/(2pq)}.$$

Hence, for each $r \geq 1$, $d_2(r) > d_1(r)$ will follow from

$$(2pq)^{\frac{r+1}{2}} \Gamma\left(\frac{r+1}{2}\right) > \frac{2}{r+1} e^{-1/(2pq)}, \quad \text{for } p \in (0, 1) \text{ and } q = 1 - p. \quad (4.12)$$

To show (4.12), set $\theta := (r+1)/2$ and

$$f(x) := x^{-\theta} e^x \Gamma(\theta + 1), \quad x := \frac{1}{2pq} \geq 2.$$

Then (4.12) is equivalent to $f(x) > 1$, for all $x \geq 2$. Since

$$f'(x) = \Gamma(\theta + 1) \left(-\theta x^{-\theta-1} e^x + x^{-\theta} e^x\right) = x^{-\theta} e^x \Gamma(\theta + 1) \left(1 - \frac{\theta}{x}\right)$$

has a unique zeros at $x = \theta$, it is sufficient to show that

$$f(\theta) = \theta^{-\theta} e^\theta \Gamma(\theta + 1) = \theta^{1-\theta} e^\theta \Gamma(\theta) > 1, \quad \text{for any } \theta \geq 1.$$

But, this follows from a classical inequality for the gamma function (cf. [18, Theorem 1])

$$\frac{\Gamma(x)}{\Gamma(y)} \geq \frac{x^{x-1} e^y}{y^{y-1} e^x}, \quad \text{for any } x \geq y \geq 1.$$

Therefore, $d_2(r) > d_1(r)$, for any $r \geq 1$. That is, with the extended \mathcal{U}_n , Theorem 3.4 provides larger constants than the ones obtained with the standard \mathcal{U}_n .

Example 4.6. For $\Phi_n(x) = e^{sx/\sqrt{2n}}$ with $s > 0$, then

$$\begin{aligned} \sum_{k: |k-2nq| \leq (2n)^\beta} \Phi_n(\delta|k-2nq|) \phi_n(k) &= \frac{1}{\sqrt{2\pi pq}} \frac{1}{\sqrt{2n}} \sum_{k: |k-2nq| \leq (2n)^\beta} \exp\left(s\delta|j_k(n)| - \frac{(j_k(n))^2}{2pq}\right) \\ &\rightarrow \frac{1}{\sqrt{2\pi pq}} \int_{-\infty}^{\infty} \exp\left(s\delta|x| - \frac{x^2}{2pq}\right) dx, \quad n \rightarrow \infty. \end{aligned} \quad (4.13)$$

Hence, (4.11) leads to

$$\liminf_{n \rightarrow \infty} \mathbb{E} \left(\exp \left(s \frac{|L_n(\mathbf{Z}_n) - \mu_n|}{\sqrt{2n}} \right) \right) \geq \mathbb{E} \left(e^{s\delta|\xi|} \right),$$

where again $\xi \sim \mathcal{N}(0, pq)$. Therefore, for any $s > 0$,

$$\liminf_{n \rightarrow \infty} \mathbb{E} \left(e^{s|V_n|} \right) \geq \mathbb{E} \left(e^{s\epsilon_0|\xi|/\sqrt{2}} \right). \quad (4.14)$$

Unlike the lower bound obtained in (4.6), the present lower bound approaches one as $s \downarrow 0$. Thus, again, the bound obtained via Theorem 3.4, with the extended \mathcal{U}_n , outperforms the one obtained with the standard \mathcal{U}_n .

Remark 4.7. Since d_2 is given by the r -th moment of $|\xi|$, clearly

$$1 + \sum_{r=1}^{\infty} d_2(r) \frac{s^r}{r!} = \mathbb{E} \left(e^{s\epsilon_0|\xi|/\sqrt{2}} \right),$$

and so in this case, the lower bound on the moment generating function obtained via the series expansion gives the same result as (4.14).

Remark 4.8. There is another way to get (4.13). By (4.9), it is enough to prove that, as $n \rightarrow \infty$,

$$\sum_{k: |k-2nq| \leq (2n)^\beta} \Phi_n(\delta|k-2nq|) \mathbb{P}(U_n = k) = \sum_{k: |k-2nq| \leq (2n)^\beta} \exp \left\{ \frac{s\delta}{\sqrt{2n}} |k-2nq| \right\} \mathbb{P}(U_n = k) \rightarrow \mathbb{E} \left(e^{s\delta|\xi|} \right).$$

To see this, note that by the CLT and the continuous mapping theorem, for any $s \in \mathbb{R}$,

$$\exp \left(\frac{s\delta}{\sqrt{2n}} |U_n - 2nq| \right) \xrightarrow{\mathcal{D}} e^{s\delta|\xi|}, \quad n \rightarrow \infty, \quad (4.15)$$

where “ $\xrightarrow{\mathcal{D}}$ ” stands for convergence in law. Moreover, for any $\epsilon > 0$, by Hoeffding’s exponential inequality,

$$\begin{aligned} \mathbb{E} \left(\exp \left(\frac{s\delta(1+\epsilon)}{\sqrt{2n}} |U_n - 2nq| \right) \right) &= 1 + \int_1^\infty \mathbb{P} \left(\exp \left(s\delta(1+\epsilon)\sqrt{2n} \left| \frac{U_n - 2nq}{2n} \right| \right) \geq x \right) dx \\ &= 1 + \int_1^\infty \mathbb{P} \left(\left| \frac{U_n}{2n} - q \right| \geq \frac{\ln x}{s\delta(1+\epsilon)\sqrt{2n}} \right) dx \\ &\leq 1 + 2 \int_1^\infty \exp \left(-4n \frac{(\ln x)^2}{2s^2\delta^2(1+\epsilon)^2n} \right) dx \\ &= 1 + 2 \int_0^\infty \exp \left(-\frac{2y^2}{s^2\delta^2(1+\epsilon)^2} + y \right) dy < \infty, \end{aligned}$$

which implies that the family of non-negative random variables

$$\left\{ \exp \left(\frac{s\delta}{\sqrt{2n}} |U_n - 2nq| \right) : n \in \mathbb{N} \right\}$$

is uniformly integrable. Let $A_n := \{U_n \notin \mathcal{U}_n\}$. Again, Hoeffding's exponential inequality leads to

$$\mathbb{P}(A_n) = \mathbb{P}(|U_n - 2nq| > (2n)^\beta) \leq 2 \exp(-2(2n)^{2\beta-1}) \rightarrow 0, \quad n \rightarrow \infty, \quad (4.16)$$

since $2\beta > 1$. Therefore, for any $s \in \mathbb{R}$,

$$\begin{aligned} & \lim_{n \rightarrow \infty} \sum_{k: |k-2nq| \leq (2n)^\beta} \exp \left(\frac{s\delta}{\sqrt{2n}} |k - 2nq| \right) \mathbb{P}(U_n = k) \\ &= \lim_{n \rightarrow \infty} \mathbb{E} \left(\mathbf{1}_{A_n^c} \exp \left(\frac{s\delta}{\sqrt{2n}} |U_n - 2nq| \right) \right) = \lim_{n \rightarrow \infty} \mathbb{E} \left(\exp \left(\frac{s\delta}{\sqrt{2n}} |U_n - 2nq| \right) \right) = \mathbb{E} \left(e^{s\delta|\xi|} \right), \end{aligned}$$

where the last equality follows from (4.15) and the uniform integrability.

4.2 Uniform Approximation

We next show that in the binary LCS case, the uniform approximation approach (Theorem 3.6) also applies provided that \mathcal{U}_n is standard. Indeed, with the standard \mathcal{U}_n , we have $m(n) = 2\sqrt{2n}$ (ignoring the rounding), and $\varphi(n) = 1/(b\sqrt{n})$, so that the condition $m(n) \geq c\varphi^{-1}(n) = bc\sqrt{n}$ holds with $c = 2\sqrt{2}/b$, and thus Theorem 3.6 applies. Therefore,

$$\mathbb{E}(\Phi(|L_n(\mathbf{Z}_n) - \mu_n|)) \geq \sum_{j=1}^{\frac{c}{8}\varphi^{-1}(n)} \Phi \left(\frac{\varepsilon_0}{2} \left(\frac{c}{8\varphi(n)} + j \right) \right) \varphi(n) + \frac{c}{8} \Phi \left(\frac{\varepsilon_0 c}{16\varphi(n)} \right) \geq \frac{c}{4} \Phi \left(\frac{\varepsilon_0 c}{16\varphi(n)} \right). \quad (4.17)$$

Again, let us apply this result to $|L_n(\mathbf{Z}_n) - \mu_n|$.

Example 4.9. For $\Phi(x) = |x|^r$, $r \geq 1$, the rightmost bound in (4.17) gives

$$\mathbb{E}(|L_n(\mathbf{Z}_n) - \mu_n|^r) \geq \frac{c}{4} \left(\frac{\varepsilon_0 c}{16\varphi(n)} \right)^r = \frac{\varepsilon_0^r}{2^{(5r+1)/2} b} n^{r/2}.$$

Note that

$$d_3(r) = d_3(r, p, \varepsilon_0) := \frac{\varepsilon_0^r}{2^{(5r+1)/2} b} < \frac{2^{\frac{3-r}{2}} \varepsilon_0^r}{b(r+1)} = d_1(r),$$

where $d_1(r)$ is the constant given in (4.4). This last fact is quite expected since we had a cruder approximation and a weaker assumption $\Delta_n(\varepsilon_0) \rightarrow 0$, as $n \rightarrow \infty$.

Next, applying the finer lower middle bound in (4.17) gives

$$\begin{aligned} \sum_{j=1}^{\frac{c}{8}\varphi^{-1}(n)} \Phi \left(\frac{\varepsilon_0}{2} \left(\frac{c}{8\varphi(n)} + j \right) \right) \varphi(n) + \frac{c}{8} \Phi \left(\frac{\varepsilon_0 c}{16\varphi(n)} \right) &= \frac{1}{b\sqrt{n}} \left(\frac{\varepsilon_0}{2} \right)^r \sum_{j=1}^{\sqrt{n}/(2\sqrt{2})} \left(\frac{\sqrt{n}}{2\sqrt{2}} + j \right)^r + \frac{\varepsilon_0^r n^{r/2}}{2^{(5r+3)/2} b} \\ &\geq \frac{\varepsilon_0^r}{b 2^r \sqrt{n}} \int_0^{\sqrt{n}/(2\sqrt{2})} \left(\frac{\sqrt{n}}{2\sqrt{2}} + x \right)^r dx + \frac{\varepsilon_0^r n^{r/2}}{2^{(5r+3)/2} b} \\ &= \frac{\varepsilon_0^r}{2^{(5r+3)/2} b} \left(\frac{2^{r+1} - 1}{r+1} + 1 \right) n^{r/2}. \end{aligned}$$

Once more, a better approximation leads to a larger constant

$$d_4(r) = d_4(r, p, \varepsilon_0) := b^{-1} \varepsilon_0^r 2^{-\frac{5r+3}{2}} \left(\frac{2^{r+1} - 1}{r + 1} + 1 \right) = \frac{2^{r+1} + r}{2(r + 1)} d_3(r) < d_1(r).$$

Example 4.10. For $\Phi(x) = e^{tx}$ with $t > 0$, the rightmost bound in (4.17) gives

$$\mathbb{E} \left(e^{|L_n(\mathbf{Z}_n) - \mu_n|t} \right) \geq \frac{c}{4} \Phi \left(\frac{\varepsilon_0 c}{16\varphi(n)} \right) = \frac{1}{\sqrt{2}b} e^{\varepsilon_0 t \sqrt{n}/(4\sqrt{2})}, \quad (4.18)$$

while the more refined middle bound, with $\rho_t = e^{\varepsilon_0 t/2}$, gives

$$\begin{aligned} \sum_{j=1}^{\frac{c}{8}\varphi^{-1}(n)} \Phi \left(\frac{\varepsilon_0}{2} \left(\frac{c}{8\varphi(n)} + j \right) \right) \varphi(n) + \frac{c}{8} \Phi \left(\frac{\varepsilon_0 c}{16\varphi(n)} \right) &= \frac{e^{\varepsilon_0 t \sqrt{n}/(4\sqrt{2})}}{b\sqrt{n}} \sum_{j=1}^{\sqrt{n}/(2\sqrt{2})} \rho_t^j + \frac{1}{2\sqrt{2}b} \rho_t^{\sqrt{n}/(2\sqrt{2})} \\ &= \frac{\rho_t}{b\sqrt{n}(\rho_t - 1)} \left(\rho_t^{\sqrt{n}/\sqrt{2}} - \rho_t^{\sqrt{n}/(2\sqrt{2})} \right) + \frac{\rho_t^{\sqrt{n}/(2\sqrt{2})}}{2\sqrt{2}b}. \end{aligned}$$

For every $t > 0$, the convergence to infinity (as $n \rightarrow \infty$) of this last bound is slower than that of the bound (4.5). Hence, again, the uniform bound is smaller (for large n) than the ones obtained in the previous subsection using Theorem 3.4. Since $t > 0$, for n large enough (depending on t),

$$\mathbb{E} \left(e^{|L_n(\mathbf{Z}_n) - \mu_n|t} \right) \geq \frac{1}{b} e^{\varepsilon_0 t \sqrt{n}/(4\sqrt{2})} \frac{\rho_t^{\frac{1}{4}\sqrt{n}+1}}{\rho_t - 1} + \frac{e^{\varepsilon_0 t \sqrt{n}/(4\sqrt{2})}}{2\sqrt{2}b} \geq \frac{1}{b} e^{\varepsilon_0 t \sqrt{n}/4} + \frac{e^{\varepsilon_0 t \sqrt{n}/(4\sqrt{2})}}{2\sqrt{2}b}.$$

Remark 4.11. Again, when estimating the moment generating function of $|V_n| = |L_n(\mathbf{Z}) - \mu_n|/\sqrt{n}$ via its series expansion, with the lower bound $\mathbb{E}(|V_n|^r) \geq d_3(r)$, for large n (uniformly in r),

$$\mathbb{E} \left(e^{s|V_n|} \right) \geq 1 + \frac{1}{\sqrt{2}b} \sum_{r=1}^{\infty} \frac{1}{r!} \left(\frac{\varepsilon_0}{4\sqrt{2}} s \right)^r = \frac{1}{\sqrt{2}b} e^{\varepsilon_0 s/(4\sqrt{2})} + 1 - \frac{1}{\sqrt{2}b},$$

which is the same bound as in (4.18) with $t = s/\sqrt{n}$, except for the constant term.

Remark 4.12. With the extended \mathcal{U}_n , the uniform approximation (Theorem 3.6) cannot be applied, since there does not exist a constant $c > 0$ such that $m(n) = 2(2n)^\beta \geq c\varphi^{-1}(n)$, for any $n \in \mathbb{N}$, with $\varphi(n)$ given in (4.10).

Remark 4.13. We have obtained several constants $d_3(r) < d_4(r) < d_1(r) < d_2(r)$ involved in lower bounding the moment $\mathbb{E}(|V_n|^r)$. The best constant, $d_2(r)$, is obtained under fast convergence with extended \mathcal{U}_n . Clearly, in order to deduce the existence of $c(r)$ such that $\mathbb{E}(|V_n|^r) \geq c(r)$, it suffices to tackle the case $r = 1$. However, it is easy to verify that in all the cases considered above (i.e., $c(r)$ being either $d_1(r)$, $d_2(r)$, $d_3(r)$ or $d_4(r)$), the constant $(c(1))^r$ is smaller than $c(r)$. Moreover, in all these cases, it is also true that $c(r+1) \geq (c(r))^{(r+1)/r}$, so that a better constant is obtained when estimating the r -th moment directly rather than estimating a lower-order moment.

5 An Upper Bound on the Rate Function

5.1 Background and Preliminary Results

The analysis right below (before Proposition 5.1) is similar to [4, Theorem 2], but with a more general notion of score function and a slightly different definition of gap^1 . Again, let $(X_n)_{n \in \mathbb{N}}$ and

¹In [4], the scoring function takes the value 1 for matches and the penalty $-\mu$ for mismatches. Moreover, the gap in [4] is an indel in one sequence, and the gap price $-\delta$ is assumed to be negative.

$(Y_n)_{n \in \mathbb{N}}$ be two independent sequences of i.i.d. random variables, and let us consider a general scoring function $S : \mathcal{A} \times \mathcal{A} \rightarrow \mathbb{R}^+$. By subadditivity,

$$L_{n+m} \geq L_n + L(\mathbf{X}_{n+1}^{n+m}; \mathbf{Y}_{n+1}^{n+m}),$$

where $L(\mathbf{X}_{n+1}^{n+m}; \mathbf{Y}_{n+1}^{n+m})$ denotes the optimal alignments score of $\mathbf{X}_{n+1}^{n+m} := (X_{n+1}, \dots, X_{n+m})$ and $\mathbf{Y}_{n+1}^{n+m} := (Y_{n+1}, \dots, Y_{n+m})$. Hence, for any $s \geq 0$,

$$\mathbb{P}(L_{n+m} \geq s(n+m)) \geq \mathbb{P}(L_n \geq sn, L(\mathbf{X}_{n+1}^{n+m}; \mathbf{Y}_{n+1}^{n+m}) \geq sm) = \mathbb{P}(L_n \geq sn) \mathbb{P}(L_m \geq sm).$$

Thus, by Fekete's lemma, for any $s \geq 0$, the following limit – the *rate function* – exists:

$$r(s) := \lim_{n \rightarrow \infty} -\frac{1}{n} \ln \mathbb{P}(L_n \geq sn) = \inf_{n \in \mathbb{N}} -\frac{1}{n} \ln \mathbb{P}(L_n \geq sn) < \infty.$$

and so,

$$\mathbb{P}(L_n \geq sn) \leq e^{-r(s)n}, \quad \text{for any } s \geq 0, n \in \mathbb{N}.$$

Since for any $n \in \mathbb{N}$,

$$\mathbb{E}(L_n) \leq \gamma^* n, \quad \text{where } \gamma^* := \lim_{n \rightarrow \infty} \frac{\mathbb{E}(L_n)}{n},$$

it follows from (2.2) that for any $s > \gamma^*$,

$$\mathbb{P}(L_n \geq sn) = \mathbb{P}\left(L_n - \mathbb{E}(L_n) \geq \left(s - \frac{\mathbb{E}(L_n)}{n}\right)n\right) \leq \mathbb{P}(L_n - \mathbb{E}(L_n) \geq (s - \gamma^*)n) \leq \exp\left(-\frac{(s - \gamma^*)^2 n}{K^2}\right).$$

Therefore, for $s > \gamma^*$,

$$r(s) \geq \frac{(s - \gamma^*)^2}{K^2}. \tag{5.1}$$

Moreover, for $s = \gamma^*$,

$$\mathbb{P}(L_n \geq \gamma^* n) = \mathbb{P}\left(L_n - \mathbb{E}(L_n) \geq \left(\gamma^* - \frac{\mathbb{E}(L_n)}{n}\right)n\right) \leq \exp\left(-\frac{n}{K^2} \left(\gamma^* - \frac{\mathbb{E}(L_n)}{n}\right)^2\right).$$

The aim of the present section is to show that the methodology developed to date allows us to partially reverse (5.1) in the LCS case. That is, we will show that there exists a universal constant $B > 0$, such that

$$r(s) \leq B(s - \gamma^*)^2, \tag{5.2}$$

for any s belonging to an upper neighborhood of γ^* . Moreover, the full claim (5.2), i.e., for all $s > \gamma^*$, is shown to hold under a uniform convergence assumption, which is not proved in general.

Hereafter and throughout this section, we will only consider the binary LCS case. For the rest of this subsection, we obtain a lower bound on $\mathbb{E}(e^{t(L_n(\mathbf{Z}_n) - \ell_n(2nq))})$ (note that there is no absolute value in the exponent), which will be used to prove the claim (5.2) in the neighborhood of γ^* . Since we are considering the special case of LCS and since we have already seen the advantage of the extended \mathcal{U}_n over the standard one, we shall use the extended \mathcal{U}_n . Throughout this section, let $p_0 > 0$ be given as in Theorem 4.1.

Proposition 5.1. *Let $t > 0$ and let $p \in (0, p_0)$. Then, there exists $\varepsilon_0 > 0$, such that*

$$\liminf_{n \rightarrow \infty} \mathbb{E} \left(\exp \left(\frac{t}{\sqrt{2n}} (L_n(\mathbf{Z}_n) - \ell_n(2nq)) \right) \right) \geq \lambda e^{pq \varepsilon_0^2 t^2 / 8},$$

where

$$\lambda = \lambda(\varepsilon_0) := \min_{t > 0} \left(1 - \mathbb{P} \left(\frac{pq \varepsilon_0 t}{2} < \xi < pqt \right) \right) \in (0, 1), \quad (5.3)$$

and where $\xi \sim \mathcal{N}(0, pq)$.

Proof. For simplicity, assume $2nq$ to be an integer. Let \mathcal{U}_n be given by (4.8). By the conditional Jensen's inequality,

$$\begin{aligned} \mathbb{E} \left(e^{t(L_n(\mathbf{Z}_n) - \ell_n(2nq))} \right) &\geq \mathbb{E} \left(e^{t(\ell_n(U_n) - \ell_n(2nq))} \right) \geq \sum_{k \in \mathcal{U}_n} e^{t(\ell_n(k) - \ell_n(2nq))} \mathbb{P}(U_n = k) \\ &= \sum_{k=2nq}^{2nq+(2n)^\beta} e^{t(\ell_n(k) - \ell_n(2nq))} \mathbb{P}(U_n = k) + \sum_{k=2nq-(2n)^\beta}^{2nq-1} e^{t(\ell_n(k) - \ell_n(2nq))} \mathbb{P}(U_n = k). \end{aligned} \quad (5.4)$$

Theorem 4.1 ensures that, with $\varepsilon_0 := \varepsilon_1 - \varepsilon_2$, the condition (3.10) holds and therefore, all the assumptions of Theorem 3.4 are satisfied. If $k \in \mathcal{U}_n$ and $k \geq 2nq$, then by Theorem 3.4, $\ell_n(k) - \ell_n(2nq) \geq \delta(k - 2nq)$, where $\delta = \varepsilon_0/2 < 1$; while if $k < 2nq$, then by (4.1), $\ell_n(k) - \ell_n(2nq) \geq k - 2nq$. Hence, by (4.9), for any $\epsilon > 0$ and n (depending on ϵ , but not on t) large enough,

$$\sum_{k=2nq}^{2nq+(2n)^\beta} \exp \left(\frac{t}{\sqrt{2n}} (\ell_n(k) - \ell_n(2nq)) \right) \mathbb{P}(U_n = k) \geq (1 - \epsilon) \sum_{k=2nq}^{2nq+(2n)^\beta} \exp \left(\frac{t\delta}{\sqrt{2n}} (k - 2nq) \right) \phi_n(k),$$

and

$$\sum_{k=2nq}^{2nq+(2n)^\beta} \exp \left(\frac{t\delta}{\sqrt{2n}} (k - 2nq) \right) \phi_n(k) \rightarrow \frac{1}{2} \left(1 + \operatorname{erf} \left(\frac{\sqrt{pq} \delta t}{\sqrt{2}} \right) \right) e^{pq \varepsilon_0^2 t^2 / 8}.$$

Similarly, for n (depending on ϵ , but not on t) large enough,

$$\sum_{k=2nq-(2n)^\beta}^{2nq-1} \exp \left(\frac{t}{\sqrt{2n}} (\ell_n(k) - \ell_n(2nq)) \right) \mathbb{P}(U_n = k) \geq (1 - \epsilon) \sum_{k=2nq-(2n)^\beta}^{2nq-1} \exp \left(\frac{t}{\sqrt{2n}} (k - 2nq) \right) \phi_n(k),$$

and

$$\sum_{k=2nq-(2n)^\beta}^{2nq-1} \exp \left(\frac{t}{\sqrt{2n}} (k - 2nq) \right) \phi_n(k) \rightarrow \frac{1}{2} \left(1 - \operatorname{erf} \left(\frac{\sqrt{pq} t}{\sqrt{2}} \right) \right) e^{pq t^2 / 2}.$$

Since $\epsilon > 0$ is arbitrarily chosen,

$$\begin{aligned} \liminf_{n \rightarrow \infty} \mathbb{E} \left(\exp \left(\frac{t}{\sqrt{2n}} (L_n(\mathbf{Z}_n) - \ell_n(2nq)) \right) \right) &\geq \frac{1}{2} \left(1 + \operatorname{erf} \left(\frac{\sqrt{pq} \delta t}{\sqrt{2}} \right) \right) e^{pq \varepsilon_0^2 t^2 / 8} + \frac{1}{2} \left(1 - \operatorname{erf} \left(\frac{\sqrt{pq} t}{\sqrt{2}} \right) \right) e^{pq t^2 / 2} \\ &> \left[1 + \frac{1}{2} \left(\operatorname{erf} \left(\frac{\sqrt{pq} \delta t}{\sqrt{2}} \right) - \operatorname{erf} \left(\frac{\sqrt{pq} t}{\sqrt{2}} \right) \right) \right] e^{pq \varepsilon_0^2 t^2 / 8} \\ &= (1 - \mathbb{P}(pq \delta t < \xi < pqt)) e^{pq \varepsilon_0^2 t^2 / 8}, \end{aligned}$$

where $\xi \sim \mathcal{N}(0, pq)$. □

To conclude this subsection, we provide an estimate on $|\mathbb{E}(L_n(\mathbf{Z}_n)) - \ell_n(2nq)|/\sqrt{2n}$, which will be useful in the sequel.

Corollary 5.2. *In the binary LCS setting, with $p \in (0, p_0)$,*

$$\limsup_{n \rightarrow \infty} \frac{|\mathbb{E}(L_n(\mathbf{Z}_n)) - \ell_n(2nq)|}{\sqrt{2n}} \leq \sqrt{\frac{2pq}{\pi}}.$$

In particular,

$$\lim_{n \rightarrow \infty} \frac{\ell_n(2nq)}{n} = \lim_{n \rightarrow \infty} \frac{\mathbb{E}(L_n)}{n} = \gamma^*. \quad (5.5)$$

Proof. By (4.1),

$$|\ell_n(k) - \ell_n(2nq)| \leq |k - 2nq|, \quad k \in \mathcal{U}_n.$$

Hence, for any $\epsilon > 0$ and n large enough,

$$\sum_{k \in \mathcal{U}_n} \frac{|\ell_n(k) - \ell_n(2nq)|}{\sqrt{2n}} \mathbb{P}(U_n = k) \leq (1 + \epsilon) \sum_{k \in \mathcal{U}_n} \frac{|k - 2nq|}{\sqrt{2n}} \phi_n(k).$$

Again,

$$\sum_{k \in \mathcal{U}_n} \frac{|k - 2nq|}{\sqrt{2n}} \phi_n(k) \rightarrow \int_{-\infty}^{\infty} \frac{|x|}{\sqrt{2\pi pq}} e^{-x^2/(2pq)} dx = \sqrt{\frac{2pq}{\pi}},$$

and by (4.16),

$$\sum_{k \notin \mathcal{U}_n} \frac{|\ell_n(k) - \ell_n(2nq)|}{\sqrt{2n}} \mathbb{P}(U_n = k) \leq \frac{n}{\sqrt{2n}} \mathbb{P}(U_n \notin \mathcal{U}_n) \leq \sqrt{2n} \exp[-2(2n)^{2\beta-1}] \rightarrow 0, \quad n \rightarrow \infty.$$

Thus,

$$\limsup_{n \rightarrow \infty} \frac{|\mathbb{E}(L_n(\mathbf{Z}_n)) - \ell_n(2nq)|}{\sqrt{2n}} \leq \limsup_{n \rightarrow \infty} \frac{\mathbb{E}(|\ell_n(U_n) - \ell_n(2nq)|)}{\sqrt{2n}} \leq \sqrt{\frac{2pq}{\pi}},$$

which clearly implies (5.5). \square

5.2 An Upper Bound on the Rate Function in the Neighborhood of γ^*

The goal of this subsection is to prove the claim (5.2) in the neighborhood of γ^* , i.e., for some constant $B > 0$ and $\tilde{\gamma} > 0$,

$$r(s) \leq B(s - \gamma^*)^2, \quad \text{for any } s \in (\gamma^*, \gamma^* + \tilde{\gamma}]. \quad (5.6)$$

Let

$$\Lambda_n(t) := \log \mathbb{E} \left(e^{tL_n(\mathbf{Z}_n)} \right), \quad t \in \mathbb{R}.$$

The following result of [14] (see also [13, Theorem 1]) will be useful in the sequel of the proof: there exists $\Lambda(t)$, $t \in \mathbb{R}$, such that

$$\lim_{n \rightarrow \infty} \frac{\Lambda_n(t)}{n} = \Lambda(t), \quad \text{for any } t \in \mathbb{R}.$$

Moreover, $r(s)$ and $\Lambda(t)$ are convex functions and are related via

$$r(s) = \Lambda^*(s) := \sup_{t \geq 0} (ts - \Lambda(t)), \quad \Lambda(t) = r^*(t) = \sup_{s \in \mathbb{R}} (ts - r(s)), \quad \text{for all } s \in \mathbb{R} \text{ and } t \geq 0. \quad (5.7)$$

5.2.1 A Global Upper Bound Under a Uniform Lower Bound

Recall from Proposition 5.1 that

$$\liminf_{n \rightarrow \infty} \mathbb{E} \left[\exp \left(\frac{t}{\sqrt{2n}} (L_n(\mathbf{Z}) - \ell_n(2nq)) \right) \right] \geq \lambda e^{pq\varepsilon_0^2 t^2/8},$$

where λ is given by (5.3). Hence, for every $0 < \lambda_1 < \lambda$ and $t > 0$,

$$\mathbb{E} \left[\exp \left(\frac{t}{\sqrt{2n}} (L_n(\mathbf{Z}_n) - \ell_n(2nq)) \right) \right] \geq \lambda_1 e^{pq\varepsilon_0^2 t^2/8}, \quad (5.8)$$

provided that n (which depends on t) is large enough .

Proposition 5.3. *Let (5.8) hold uniformly over $[0, \infty)$, i.e., let there exists $n_0 \in \mathbb{N}$ (independent of t) such that, whenever $n \geq n_0$, (5.8) holds true for every $t \geq 0$, then (5.2) holds true for all $s > \gamma^*$.*

Proof. Let $\lambda_2 := pq\varepsilon_0^2/8$. Since (5.8) holds uniformly in t , for $n \geq n_0$,

$$\Lambda_n \left(\frac{t}{\sqrt{2n}} \right) \geq \frac{\ell_n(2nq)t}{\sqrt{2n}} + \ln \lambda_1 + \lambda_2 t^2, \quad \text{for any } t \geq 0.$$

With $u = t/\sqrt{2n}$, for $n \geq n_0$,

$$\frac{\Lambda_n(u)}{n} \geq \frac{\ell_n(2nq)u}{n} + \frac{\ln \lambda_1}{n} + 2\lambda_2 u^2, \quad \text{for any } u \geq 0,$$

and thus (recalling (5.5)),

$$\Lambda(u) = \lim_{n \rightarrow \infty} \frac{\Lambda_n(u)}{n} \geq \gamma^* u + 2\lambda_2 u^2, \quad \text{for any } u \geq 0.$$

Now, for every $s > \gamma^*$, $u \geq 0$,

$$su - \Lambda(u) \leq (s - \gamma^*) u - 2\lambda_2 u^2,$$

which implies that,

$$r(s) = \sup_{u \geq 0} (su - \Lambda(u)) \leq \sup_{u \geq 0} [(s - \gamma^*) u - 2\lambda_2 u^2] = \frac{(s - \gamma^*)^2}{8\lambda_2}.$$

Therefore, (5.2) holds true, for all $s > \gamma^*$, with $B = 1/(8\lambda_2) = 1/(4pq\delta^2)$. \square

5.2.2 Binomial Approximation With the Further Extended \mathcal{U}_n

In this part, we derive (5.6) without assuming that (5.8) holds uniformly in t . Let $M_q(t)$ and $K_q(t)$ be respectively the moment generating function and the cumulant generating function of the Rademacher law with parameter q , i.e.,

$$M_q(t) := e^{t(1-q)}q + e^{-tq}(1-q), \quad K_q(t) := \ln M_q(t) = \ln \left(e^{t(1-q)}q + e^{-tq}(1-q) \right).$$

We start with the following general theorem.

Theorem 5.4. *Let $p \in (0, p_0)$. Let there exist $t_0 > 0$ and $\delta > 0$ such that, for any $t \in [0, t_0]$, whenever n is large enough (possibly depending on t),*

$$\mathbb{E} \left(e^{t(L_n(\mathbf{Z}_n) - \ell_n(2nq))} \right) \geq \nu_n M_q^{2n}(\delta t), \quad (5.9)$$

where $(\nu_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$ is such that $\ln \nu_n = o(n)$, as $n \rightarrow \infty$. Then, there exist constants $B > 0$ and $\tilde{\gamma} > 0$ such that (5.6) holds true.

Proof. Let $t \in [0, t_0]$. Taking logarithms in (5.9) leads to

$$\Lambda_n(t) = \ln \mathbb{E} \left(e^{tL_n(\mathbf{Z}_n)} \right) \geq t \ell_n(2nq) + 2n K_q(\delta t).$$

Dividing both sides of the above equality by n , and using (5.5), lead to

$$\Lambda(t) \geq t\gamma^* + 2K_q(\delta t). \quad (5.10)$$

Now for any $s \in \mathbb{R}$,

$$ts - \Lambda(t) \leq ts - t\gamma^* - 2K_q(\delta t) = t(s - \gamma^*) - 2K_q(\delta t),$$

which implies that

$$\sup_{t \in [0, t_0]} (ts - \Lambda(t)) \leq \sup_{t \in [0, t_0]} [t(s - \gamma^*) - 2K_q(\delta t)]. \quad (5.11)$$

We next show that $\Lambda'(0+) = \gamma^*$ (recall that $\Lambda(t)$ and $r(s)$ are related to each other only for $t \in [0, \infty)$, see (5.7)). First, by (5.10), since $K_q(\delta t) \geq 0$ for any $t \geq 0$, we have

$$\Lambda(t) \geq \gamma^* t, \quad \text{for any } t \geq 0. \quad (5.12)$$

Next, by definition, $r(s) = 0$ for any $s < \gamma^*$. As stated in [13, Theorem 1], r is a convex and, therefore, continuous function, and so $r(\gamma^*) = 0$. Thus, the function r is equal to zero up to γ^* and is strictly increasing afterwards, and so,

$$\Lambda(t) = \sup_{s \in \mathbb{R}} (ts - r(s)) = \sup_{s > \gamma^*} (ts - r(s)).$$

Using (5.1) with $K = 1$ in the LCS case, for any $t > 0$,

$$\Lambda(t) = \sup_{s > \gamma^*} (ts - r(s)) \leq \sup_{s > \gamma^*} \left[ts - (s - \gamma^*)^2 \right] = ts_t - (s_t - \gamma^*)^2 = t\gamma^* + \frac{1}{4}t^2, \quad (5.13)$$

where $s_t := t/2 + \gamma^* > \gamma^*$, $t > 0$. Combining (5.12) and (5.13), it follows that $\Lambda'(0+) = \gamma^*$, which, together with (5.11), implies that there exists $\tilde{\gamma}_1 > 0$, such that for any $s \in (\gamma^*, \gamma^* + \tilde{\gamma}_1]$,

$$r(s) = \sup_{t \geq 0} (ts - \Lambda(t)) = \sup_{t \in [0, t_0]} (ts - \Lambda(t)) \leq \sup_{t \in [0, t_0]} [t(s - \gamma^*) - 2K_q(\delta t)].$$

Now,

$$\begin{aligned} \sup_{t \in [0, t_0]} [t(s - \gamma^*) - 2K_q(\delta t)] &= 2 \sup_{t \in [0, t_0]} \left[\delta t \left(q + \frac{s - \gamma^*}{2\delta} \right) - \ln \left(qe^{t\delta} + (1 - q) \right) \right] \\ &= 2 \sup_{u \in [0, \delta t_0]} \left[u \left(q + \frac{s - \gamma^*}{2\delta} \right) - \ln (qe^u + (1 - q)) \right]. \end{aligned}$$

With

$$x := q + \frac{s - \gamma^*}{2\delta},$$

we obtain that the solution of

$$x = \frac{qe^u}{qe^u + (1 - q)}$$

is

$$u(x) = \ln \frac{x(1 - q)}{q(1 - x)}.$$

Clearly, there exists $\tilde{\gamma}_2 > 0$, such that when $s - \gamma^* \leq \tilde{\gamma}_2$, $u(x) \in [0, \delta t_0]$, and in this case

$$\begin{aligned} \sup_{u \in [0, \delta t_0]} \left[u \left(q + \frac{s - \gamma^*}{2\delta} \right) - \ln(qe^u + (1 - q)) \right] &= \sup_{u \geq 0} \left[u \left(q + \frac{s - \gamma^*}{2\delta} \right) - \ln(qe^u + (1 - q)) \right] \\ &= u(x)x - \ln(qe^{u(x)} + (1 - q)) \\ &= x \ln \frac{x}{q} + (1 - x) \ln \frac{1 - x}{1 - q} \\ &=: D(x||q). \end{aligned}$$

Since for $w > 0$, $\ln(1 + w) < w$, and since for $w \in (0, 1)$, $\ln(1 - w) < -w$, it follows that

$$D(x||q) \leq x \frac{x - q}{q} - (1 - x) \frac{x - q}{1 - q} = \frac{(x - q)^2}{q(1 - q)} = \frac{1}{4\delta^2 q(1 - q)} (s - \gamma^*)^2.$$

Therefore, (5.6) holds with $B = 1/(4\delta^2 q(1 - q))$ and $\tilde{\gamma} = \min(\tilde{\gamma}_1, \tilde{\gamma}_2)$. \square

Remark 5.5. By examining the above proof, it is easy to see that if (5.9) holds for any $t > 0$, then (5.2) holds for any $s > \gamma^*$.

Let us return to (5.9) and examine under which conditions the hypothesis of Theorem 5.4 are satisfied. Since $U_n \sim B(2n, q)$,

$$\sum_{k=0}^{2n} e^{t\delta(k-2nq)} \mathbb{P}(U_n = k) = M_q^{2n}(\delta t),$$

one might try to use the first term in (5.4) to get (5.9):

$$\mathbb{E} \left(e^{t(L_n(\mathbf{Z}_n) - \ell_n(2nq))} \right) \geq \sum_{k \in \mathcal{U}_n, k \geq 2nq} e^{t(\ell_n(k) - \ell_n(2nq))} \mathbb{P}(U_n = k).$$

Thus (5.9) holds if,

$$\sum_{k=2nq}^{2nq+(2n)^\beta} e^{t\delta(k-2nq)} \mathbb{P}(U_n = k) \geq \nu_n M_q^{2n}(\delta t),$$

for a sequence $(\nu_n)_{n \in \mathbb{N}}$ of positive reals such that $\ln \nu_n = o(n)$ (as $n \rightarrow \infty$). Unfortunately, there does not exist such $(\nu_n)_{n \in \mathbb{N}}$, since

$$\frac{1}{n} \ln \left(\sum_{k=2nq}^{2nq+(2n)^\beta} e^{t\delta(k-2nq)} \mathbb{P}(U_n = k) \right) < \frac{1}{n} \ln \left(\exp \left(t\delta(2n)^\beta \right) \right) \rightarrow 0, \quad n \rightarrow \infty.$$

To show a similar requirement, we need to further enlarge \mathcal{U}_n . The following lemma shows how to do so.

Lemma 5.6. *For any $\eta > 0$, there exists $b := b(\eta) > 0$ such that for every $n \in \mathbb{N}$ large enough,*

$$\inf_{k: |k-2nq| < bn} \mathbb{P}(U_n = k) \geq e^{-\eta n}.$$

Proof. Since $U_n \sim B(2n, q)$,

$$\mathbb{P}(U_n = k) = \binom{2n}{k} q^k (1-q)^{2n-k}, \quad k = 0, \dots, 2n.$$

Next (cf. [9, Example 12.1.3]),

$$\binom{2n}{k} \geq \frac{1}{2n+1} \exp \left(2n h_e \left(\frac{k}{2n} \right) \right), \quad k = 0, \dots, 2n, \quad n \in \mathbb{N},$$

where h_e is the binary entropy function of base e :

$$h_e(q) = -q \ln q - (1-q) \ln(1-q), \quad q \in (0, 1).$$

Hence,

$$\mathbb{P}(U_n = k) \geq \frac{1}{2n+1} \exp \left(2n \left[h_e \left(\frac{k}{2n} \right) + \frac{k}{2n} \ln q + \left(1 - \frac{k}{2n} \right) \ln(1-q) \right] \right).$$

The continuous function $g(x) = h_e(x) + x \ln q + (1-x) \ln(1-q)$, $x \in (0, 1)$, is such that $g(x) \leq 0$, for any $x \in (0, 1)$, with $g(q) = 0$. Thus, for any $\eta > 0$, there exists $b(\eta) > 0$ such that, whenever $|x - q| \leq b/2$, then $g(x) \geq -\eta/4$, and so, if $|k - 2nq| \leq b(\eta)n$,

$$h_e \left(\frac{k}{2n} \right) + \frac{k}{2n} \ln q + \left(1 - \frac{k}{2n} \right) \ln(1-q) > -\frac{\eta}{4}.$$

Therefore, for n (depending only on η) large enough,

$$\inf_{k: |k-2nq| < bn} \mathbb{P}(U_n = k) \geq \frac{1}{2n+1} \exp \left[2n h_e \left(\frac{k}{2n} \right) \right] q^k (1-q)^{2n-k} \geq \frac{1}{2n+1} \exp \left(-\frac{\eta}{2} n \right) \geq e^{-\eta n}.$$

The proof is now complete. \square

Using Lemma 5.6, we can now verify the condition (5.9) of Theorem 5.4.

Theorem 5.7. *Let $p \in (0, p_0)$. Let $\varepsilon_0 := \varepsilon_1 - \varepsilon_2 > 0$, where $\varepsilon_1 > 0$ and $\varepsilon_2 > 0$ are given in Theorem 4.1, and let $\delta := \varepsilon_0/2$. Let $t_0 > 0$ be the unique positive solution to*

$$2\delta t - b^2 - 2 \ln M_q(\delta t) = 0. \tag{5.14}$$

Then, for any $\epsilon > 0$ and any $t \in [0, t_0]$, there exists $N(t) \in \mathbb{N}$ such that for any $n \geq N(t)$, (5.9) holds true with $\nu_n \equiv 1 - \epsilon$, $n \in \mathbb{N}$.

Proof. By Lemma 5.6, the set \mathcal{U}_n can now be further enlarged. Indeed, taking $0 < \eta < c_1$ (where $c_1 > 0$ is given in Theorem 4.1), there exists $b := b(\eta) > 0$ such that

$$\sup_{k: |k-2nq| \leq bn} \frac{e^{-c_1 n}}{\mathbb{P}(U_n = k)} \leq e^{(\eta - c_1)n} \rightarrow 0, \quad n \rightarrow \infty.$$

Let us therefore take

$$\mathcal{U}_n = [2nq - bn, 2nq + bn] \cap \mathcal{S}_n.$$

With $\varphi(n) = e^{-\eta n}$ and by Theorem 4.1, the inequalities (3.10) (with $\varepsilon_0 = \varepsilon_1 - \varepsilon_2$) and (3.11) are both satisfied. Since in our LCS case (3.8) also holds, all the assumptions of Theorem 3.4 are verified and thus, for any $k \in \mathcal{U}_n$, $\ell_n(k+1) - \ell_n(k) \geq \delta = \varepsilon_0/2$, provided that n is large enough. Therefore, by the conditional Jensen's inequality,

$$\begin{aligned} \mathbb{E} \left(e^{t(L_n(\mathbf{Z}_n) - \ell_n(2nq))} \right) &\geq \mathbb{E} \left(e^{t(\ell_n(U_n) - \ell_n(2nq))} \right) \geq \sum_{k=2nq}^{2nq+bn} e^{t(\ell_n(k) - \ell_n(2nq))} \mathbb{P}(U_n = k) \\ &\geq \sum_{k=2nq}^{2nq+bn} e^{t\delta(k-2nq)} \mathbb{P}(U_n = k) =: S_n^{(3)}(t). \end{aligned} \quad (5.15)$$

On the other hand,

$$\begin{aligned} S_n(t) &:= M_q(\delta t)^{2n} = \mathbb{E} \left(e^{\delta t(U_n - 2nq)} \right) = \left[e^{\delta t(1-q)} q + e^{-\delta t q} (1-q) \right]^{2n} \\ &= \sum_{k=0}^{2nq-bn-1} e^{\delta t(k-2nq)} \mathbb{P}(U_n = k) + \sum_{k=2nq-bn}^{2nq-1} e^{\delta t(k-2nq)} \mathbb{P}(U_n = k) \\ &\quad + \sum_{k=2nq}^{2nq+bn} e^{\delta t(k-2nq)} \mathbb{P}(U_n = k) + \sum_{k=2nq+bn+1}^{2n} e^{\delta t(k-2nq)} \mathbb{P}(U_n = k) \\ &=: S_n^{(1)}(t) + S_n^{(2)}(t) + S_n^{(3)}(t) + S_n^{(4)}(t). \end{aligned}$$

It is easy to see that for every $t > 0$,

$$S_n^{(2)}(t) \leq \sum_{k=2nq-bn}^{2nq-1} e^{t\delta(k-2nq)} = \sum_{j=1}^{nb} e^{-t\delta j} = \frac{e^{-t\delta} - e^{-t\delta(nb)}}{1 - e^{-t\delta}} \rightarrow \frac{1}{e^{t\delta} - 1} =: c(t), \quad n \rightarrow \infty.$$

Moreover, for any $t > 0$,

$$S_n^{(1)}(t) + S_n^{(4)}(t) = \left(\sum_{k=0}^{2nq-bn-1} + \sum_{k=2nq+bn+1}^{2n} \right) e^{t\delta(k-2nq)} \mathbb{P}(U_n = k) < e^{2\delta t n} \mathbb{P}(|U_n - 2nq| > nb) \leq 2e^{(2\delta t - b^2)n},$$

where the last inequality follows from Hoeffding's exponential inequality. Thus, for $t > 0$,

$$1 = \frac{S_n^{(1)}(t) + S_n^{(2)}(t) + S_n^{(3)}(t) + S_n^{(4)}(t)}{S_n(t)} \leq \frac{c(t)}{S_n(t)} + \frac{S_n^{(3)}(t)}{S_n(t)} + 2 \exp \left((2\delta t - b^2 - 2 \ln M_q(\delta t)) n \right).$$

Hence, for any $0 < t < t_0$, where t_0 is the solution to (5.14),

$$\frac{S_n^{(3)}(t)}{S_n(t)} \rightarrow 1, \quad \text{as } n \rightarrow \infty.$$

Together with (5.15), it follows that, for every $t \in (0, t_0]$, there exists $N(t) \in \mathbb{N}$, such that for any $n \geq N(t)$,

$$\mathbb{E} \left(e^{t(L_n(\mathbf{Z}_n) - \ell_n(2nq))} \right) \geq S_n(t) \geq (1 - \epsilon) M_q(\delta t)^{2n},$$

which completes the proof. \square

6 Beyond the Binary LCS Case: Discussion

In this paper we applied the general methodology developed in Section 3 to the binary LCS case with strongly asymmetric distributions. From Theorem 4.1, this setup is one of few known models where the existence of a suitable random transformation and of ε_0 (such that (3.8) and $\Delta_n(\varepsilon_0) \rightarrow 0$ both hold) have been proven to exist. In the binary LCS case, Theorem 4.1 assumes a very asymmetric distribution, and further research is needed to relax this asymmetry assumption.

Other setups where the assumptions of Theorem 3.4 have been proved to hold were studied in [17], [21] and [22]. For example, in [21] and [22], the random variables $(X_n)_{n \in \mathbb{N}}$ and $(Y_n)_{n \in \mathbb{N}}$ are all i.i.d., every letter $\beta \in \mathcal{A}$ is taken with a positive probability (i.e., $\mathbb{P}(X_1 = \beta) > 0$), and the scoring function S as well as the distribution of X_1 do satisfy the following *asymmetry assumption*.

Assumption 6.1. *There exist $a, b \in \mathcal{A}$ such that*

$$\sum_{\beta \in \mathcal{A}} \mathbb{P}(X_1 = \beta) (S(b, \beta) - S(a, \beta)) > 0. \quad (6.1)$$

For the binary alphabet $\mathcal{A} = \{a, b\}$, condition (6.1) reads

$$(S(b, a) - S(a, a)) \mathbb{P}(X_1 = a) + (S(b, b) - S(b, a)) \mathbb{P}(X_1 = b) > 0.$$

Since S is symmetric and one could exchange a and b , the condition (6.1) actually becomes

$$(S(b, a) - S(a, a)) \mathbb{P}(X_1 = a) + (S(b, b) - S(b, a)) \mathbb{P}(X_1 = b) \neq 0.$$

When $S(b, b) = S(a, a) > S(b, a)$ (recall that S is assumed to be symmetric and non-constant), then Assumption 6.1 is satisfied if and only if $\mathbb{P}(X_1 = a) \neq \mathbb{P}(X_1 = b)$. Thus, in terms of the distribution of X_i and of the scoring function S , (6.1) requires much less than Theorem 4.1. However, the price to be paid there is in terms of δ . Namely, the analogue of Theorem 4.1 can be shown to hold only if the gap penalty $-\delta$ is sufficiently high. The theorem itself is as follows (cf. [21, Theorem 3.1]):

Theorem 6.2. *Let Assumption 6.1 hold. Let $a, b \in \mathcal{A}$ be as in (6.1), and let the random transformation \mathcal{R} turn the letter a in \mathbf{Z}_n into the letter b . Then there exist constants $\delta_0 < 0$, $\epsilon_0 > 0$, $\alpha > 0$, and $n_0 \in \mathbb{N}$, such that for any $\delta < \delta_0$ and $n \geq n_0$,*

$$\mathbb{P}(\mathbb{E}(L_n(\mathcal{R}(\mathbf{Z}_n)) - L_n(\mathbf{Z}_n) | \mathbf{Z}_n) \geq \epsilon_0) \geq 1 - e^{-\alpha n}.$$

Remark 6.3. Typically $\delta_0 < 0$, so that the condition $\delta < \delta_0$ indicates that the gap penalty $-\delta$ has to be sufficiently large. Hence, the result does not apply to the LCS case. Intuitively, the larger the gap penalty (the smaller the gap price), the fewer gaps appear in the optimal alignment, so that the optimal alignment is closer to the pairwise comparison (Hamming distance). Some methods for determining a sufficient δ_0 , as well as some examples, are discussed in [22]. We believe that the assumption on δ can be relaxed so that Theorem 6.2 holds under more general assumptions.

If the assumptions of Theorem 6.2 hold (i.e., the scoring function S and the law of X_1 satisfy Assumption 6.1 and δ is sufficiently small) and the alphabet $\mathcal{A} = \{a, b\}$ is binary, then the situation is exactly as in the binary LCS case considered in Section 4 except that A (see (3.9)) might not be 1. However, the constant A is not essential in any of the proofs, so all the results of Sections 4 and 5 continue to hold (with appropriate changes stemming from A). If $|\mathcal{A}| > 2$, then to apply Theorem 6.2, one should use a more general version of Theorem 3.4, see [21, Theorem 2.2]. The same theorem should apply to the LCS case when the alphabet is not binary, but the model is still strongly asymmetric, i.e., one letter has to have a very large probability.

References

- [1] K. S. Alexander, The Rate of Convergence of the Mean Length of the Longest Common Subsequence, *The Annals of Applied Probability*, Vol. 4, No. 4, pp. 1074-1082, 1994.
- [2] S. Amsalu, C. Houdré and H. Matzinger, Sparse Long Blocks and the Micro-Structure of the Longest Common Subsequences, *Journal of Statistical Physics*, Vol. 154, Issue 6, pp. 1516-1549, 2014.
- [3] S. Amsalu, C. Houdré and H. Matzinger, Sparse Long Blocks and the Variance of the Length of Longest Common Subsequences in Random Words, arXiv:1204.1009v2, 2016.
- [4] R. Arratia and M. S. Waterman, A Phase Transition for the Score in Matching Random Sequences Allowing Deletions, *The Annals of Applied Probability*, Vol. 4, No. 1, pp. 200-225, 1994.
- [5] F. Bonetto and H. Matzinger, Fluctuations of the Longest Common Subsequence in the Asymmetric Case of 2- and 3-Letter Alphabets, *Latin American Journal of Probability and Mathematical Statistics*, Vol. 2, pp. 195-216, 2006.
- [6] S. Boucheron, G. Lugosi and P. Massart, *Concentration Inequalities: A Nonasymptotic Theory of Independence*, Oxford University Press, Oxford, United Kingdom, 2013.
- [7] N. Christianini and M. W. Hahn, *Introduction to Computational Genomics: A Case Studies Approach*, Cambridge University Press, Cambridge, United Kingdom, 2007.
- [8] V. Chvátal and D. Sankoff, Longest Common Subsequences of Two Random Sequences, *Journal of Applied Probability*, Vol. 12, No. 2, pp. 306-315, 1975.
- [9] T. Cover and J. Thomas, *Elements of Information Theory*, Second Edition, John Wiley & Sons, Inc., Hoboken, New Jersey, 2006.

- [10] R. Durbin, S. Eddy, A. Krogh and G. Mitchison, *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, Cambridge University Press, Cambridge, United Kingdom, 1998.
- [11] C. Durringer, J. Lember and H. Matzinger, Deviation from the Mean in Sequence Comparison with a Periodic Sequence, *Latin American Journal of Probability and Mathematical Statistics*, Vol. 3, pp 1-29, 2007.
- [12] R. Gong, C. Houdré and Ü. Işlak, A Central Limit Theorem for the Optimal Alignments Score in Multiple Random Words, arXiv:1512.05699v2, 2016.
- [13] S. Grossmann and B. Yakir, Large Deviations for Global Maxima of Independent Superadditive Processes with Negative Drift and an Application to Optimal Sequence Alignments, *Bernoulli*, Vol. 10, No. 5, pp. 829-845, 2004.
- [14] J. M. Hammersley, Postulates for Subadditive Processes, *The Annals of Probability*, Vol. 2, No. 4, pp. 652-680, 1974.
- [15] C. Houdré and Ü. Işlak, A Central Limit Theorem for the Length of the Longest Common Subsequences in Random Words, arXiv:1408.1559v3, 2015.
- [16] C. Houdré and J. Ma, On the Order of the Central Moments of the Length of the Longest Common Subsequences in Random Words, To appear in *High Dimensional Probability VII: The Cargèse Volume*, Birkhäuser, 2016.
- [17] C. Houdré and H. Matzinger, On the Variance of the Optimal Alignments Score for Binary Random Words and an Asymmetric Scoring Function, *Journal of Statistical Physics*, Vol. 164, Issue 3, pp. 693-734, 2016.
- [18] J. D. Kečkić and P. M. Vasić, Some Inequalities for the Gamma Function, *Publications de L'institut Mathématique*, Nouvelle Série, tome 11, Vol. 25, pp. 107-114, 1971.
- [19] J. Lember and H. Matzinger, Standard Deviation of the Longest Common Subsequence, *The Annals of Probability*, Vol. 37, No. 3, pp. 1192-1235, 2009.
- [20] J. Lember, H. Matzinger and F. Torres, The Rate of the Convergence of the Mean Score in Random Sequence Comparison, *The Annals of Applied Probability*, Vol. 22, No. 3, pp. 1046-1058, 2012.
- [21] J. Lember, H. Matzinger and F. Torres, General Approach to the Fluctuations Problem in Random Sequence Comparison, arXiv:1211.5072v1, 2012.
- [22] J. Lember, H. Matzinger and F. Torres, Proportion of Gaps and Fluctuations of the Optimal Score in Random Sequence Comparison In *Limit Theorems in Probability, Statistics and Number Theory (In Honor of Friedrich Götze)*, Springer, Vol. 42, pp. 207-234, 2013.
- [23] C. Y. Lin and F. J. Och, Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram Statistics, In *ACL'04: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, pp. 605, 2004.

- [24] I. D. Melamed, Automatic Evaluation and Uniform Filter Cascades for Inducing N-Best Translation Lexicons, *In Proceedings of the Third Workshop on Very Large Corpora*, 1995.
- [25] I. D. Melamed, Bitext Maps and Alignment via Pattern Recognition, *Computational Linguistics*, Vol. 25, Issue 1, pp. 107-130, 1999.
- [26] P. A. Pevzner, *Computational Molecular Biology: An Algorithmic Approach*, MIT Press, Cambridge, 2000.
- [27] A. N. Shiryaev, *Probability*, Second Edition, Springer-Verlag, New York, 1995.
- [28] T. F. Smith and M. S. Waterman, Identification of Common Molecular Subsequences, *Journal of Molecular Biology*, Vol. 147, Issue 1, pp. 195-197, 1981.
- [29] J. M. Steele, An Efron-Stein Inequality for Nonsymmetric Statistics, *The Annals of Statistics*, Vol. 14, No. 2, pp. 753-758, 1986.
- [30] F. Torres, On the Probabilistic Longest Common Subsequence Problem for Sequences of Independent Blocks, *Ph.D Thesis*, Bielefeld University, 2009.
- [31] M. S. Waterman, Estimating Statistical Significance of Sequence Alignments, *Philosophical Transactions of the Royal Society: Biological Sciences*, Vol. 344, Issue 1310, pp. 383-390, 1994.
- [32] M. S. Waterman, *Introduction to Computational Biology: Maps, Sequences and Genomes*, Chapman & Hall / CRC Press, 1995.
- [33] C. C. Yang and K. W. Li, Automatic Construction of English/Chinese Parallel Corpora, *Journal of the American Society for Information Science and Technology*, Vol. 54, Issue 8, pp. 730-742, 2003.